2014

Analysis of Injury Severity of Drivers Involved in Single-Vehicle andTwo-Vehicle Crashes on Ontario Highways

Xuancheng Li
University of Windsor

Analysis of Injury Severity of Drivers Involved in Single-Vehicle and Two-Vehicle Crashes on Ontario Highways

By

Xuancheng Li

A Thesis
Submitted to the Faculty of Graduate Studies
through the Department of Civil and Environmental Engineering
in Partial Fulfillment of the Requirements for
the Degree of Master of Applied Science
at the University of Windsor

Windsor, Ontario, Canada

2014

Analysis of Injury Severity of Drivers Involved in Single-Vehicle and Two-Vehicle Crashes on Ontario Highways

By

Xuancheng Li

APPROVED BY:

Dr. H. Maoh,
Department of Civil and Environmental Engineering

Dr. W. Anderson,
Department of Political Science

J. Tofflemire,
Department of Civil and Environmental Engineering

Dr. C. Lee, Advisor
Department of Civil and Environmental Engineering

August 21, 2014

Declaration of Previous Publication

This thesis includes material from two original papers that have been previously published or submitted for publication in a peer reviewed journal, as follows:

Thesis Chapter	Publication title/full citation	Publication status
Chapter 5	*Lee, C., Li, X., 2014. Analysis of Injury Severity of Drivers Involved in Single- and Two-Vehicle Crashes on Highways in Ontario. Accident Analysis and Prevention, Vol. 71, pp. 286-295.*	*published*
Chapter 5	*Lee, C., Li, X., 2014. Predicting Driver's Severe Injury in Single-Vehicle and Two-Vehicle Crashes Using Boosted Regression Trees. Submitted for presentation at 94th Transportation Research Board Annual Meeting and publication in Transportation Research Record: Journal of the Transportation Research Board.*	*in press*

I certify that I have obtained a written permission from the copyright owner(s) to include the above published material(s) in my thesis. I certify that the above material describes work completed during my registration as a graduate student at the University of Windsor.

I declare that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright

Act, I certify that I have obtained written permission from the copyright owner(s) to include such material(s) in my thesis.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

Abstract

This study applies both parametric model (Heteroscedastic Ordered Logit (HOL)) and non-parametric models (Random Forest, Classification and Regression Tree (CART), and Boosted Regression Tree (BRT)) to analysis of driver's injury severity in single-vehicle and two-vehicle crashes on highways. The HOL model not only estimates quantitative effects of significant explanatory variables, but also captures heteroscedasticity (i.e. variation in the unobserved effects among observations) of the variables such as head-on collision, abnormal conditions and female drivers. On the other hand, the BRT model effectively captures nonlinear effects of continuous variables including truck percentage, AADT, driver's age and vehicle age on severe injury. It was found that the BRT model predicted driver's injury severity more accurately than the HOL and CART models for both single-vehicle and two-vehicle crashes. Based on the model results, some remedial treatments are discussed to reduce driver's injury severity in crashes on highways. It is recommended that both HOL and BRT models are used for more accurate prediction of crash injury severity.

Acknowledgments

Foremost, I would like to express my sincere gratitude to my advisor Dr. C. Lee for the continuous support of my study and research. His guidance helped me in all the time of research and writing of this thesis.

Besides my advisor, I would like to thank the rest of my thesis committee: Dr. W. Anderson, Dr. H. Maoh, Mr. J. Tofflemire and Dr. A. EI Ragaby, for their encouragement, insightful comments, and hard questions.

I would like to thank my friends with me all the time. Last but not least, I would like to thank my family: my parents and my fiancée, for supporting me no matter what.

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1. Background

Traffic crashes cause significant losses to society and may result in injury, death, vehicle damage, and property damage, etc. In 2010, the numbers of motor vehicle fatalities and serious injuries in Canada were 2,227 and 11,226, respectively (Transport Canada, 2012). In the same year, there were 579 fatalities and 2,558 serious injuries in Ontario (Ministry of Transportation Ontario, 2012). The social cost by traffic crashes including property damages/insurance claims, hospital/health care costs, first responders (police, fire, and ambulance services) and traffic delays was enormous. The social costs of motor vehicle collisions in Canada in 2004 were estimated to be $63 billion (Voden et al., 2007). In particular, according to Ministry of Transportation Ontario (2007), fatal collisions account for less than 1% of reportable collisions in Ontario in 2004, their social costs were 64% ($11 billion) of total social costs.

In addition, collisions involving trucks usually result in more severe injuries and social costs. There was an annual average of 8,985 heavy truck casualty collisions in 2010 (Transport Canada, 2010). These collisions represent 7% of all collisions, 18% of fatal collisions and 15% ($3 billion) of the social costs (Transport Canada, 2010). Higher proportion of fatal collision is mainly because collisions between light (passenger vehicles such as sedan and minivan) and heavy vehicles (pick-up trucks and heavy-duty trucks) result in more severe damages for light vehicles. Furthermore, since the impacts of collision with fixed objects on vehicles in single-vehicle crashes are different from the impacts of collision with other vehicles (mostly moving objects) in multi-vehicle crashes.

Due to this difference, injury severity has been analyzed for single-vehicle crashes and multi-vehicle crashes separately (Wang and Kockelman, 2005; Savolainen and Mannering, 2007; Chen and Chen, 2011; Weiss et al., 2014).

Injury severity is also influenced by the other factors such as seat belt usage, drinking and driving, fatigue. For example, percentage of fatality increased with the blood alcohol concentration increases (Transport Canada, 2010). Collisions involving drivers with alcohol are only 3% of all collisions but they represent 18% ($3 billion) of social costs, 24% of fatal collisions (Transport Canada, 2010). Therefore, the influence of driver conditions should be considered in the analysis of injury severity.

However, there were some limitations in the past studies. For two-vehicle crashes, most studies only considered the effects of one vehicle on driver's injury severity. However, it is expected that driver's injury severity is not only affected by characteristics of his/her own vehicle, but also characteristics of a partner vehicle. This is because sizes and weights may differ between the two vehicles and this difference has differential impacts of collision on each vehicle. Also, there is a lack of study on the comparison of injury severity between single-vehicle crashes and two-vehicle crashes to identify differential effects of explanatory variables.

1.2. Objectives of Thesis

The objectives of this research are as follows:

1) to identify the risk factors that significantly influence the injury severity of drivers involved in single-vehicle and two-vehicle crashes considering above limitations,

2) to investigate effects of these factors on injury severity using both advanced parametric and non-parametric models,

3) to evaluate the accuracy of predicted injury severity between parametric and non-parametric models, and

4) to suggest appropriate methods of reducing injury severity based on the conducted analysis.

1.3. Organization of Thesis

This thesis is organized into six chapters as follows:

- Chapter 2 reviews past studies on factors affecting crash injury severity and compares injury severity between single-vehicle and two-vehicle crashes.
- Chapter 3 describes the crash, road geometry and traffic data used for the analysis.
- Chapter 4 explains the parametric and non-parametric models which identify the relationship between driver's injury severity and explanatory variables. The chapter also discusses advantages and disadvantages of each model.
- Chapter 5 presents the results of the models and discusses the findings.
- Chapter 6 draws conclusions based on the model results and recommends future work.

2. Literature Review

2.1. Factors Affecting Injury Severity

Various studies have investigated injury severity using both parametric models and non-parametric models. Parametric models include multinomial logit (MNL) model, nested logit (NL) model, ordered logit (OL) or ordered probit (OP) model, heteroskedastic ordered logit (HOL) model, and mixed logit (MXL) model. Non-parametric models include classification and regression trees (CART), random forests method and boosted regression trees (BRT).

Using these models, researchers have analyzed the effects of many factors on injury severity and predict the potential injury levels under various conditions. In general, the factors affecting injury severities are categorized into the following four groups: 1) driver characteristics; 2) vehicle characteristics; 3) road geometric characteristics; and 4) environmental characteristics.

First, driver characteristics include driver demographic factors such as age and gender. Zhang et al. (2000) reported that older drivers are more likely to be killed or seriously injured in traffic crashes than middle-age drivers. Similarly, Kim et al. (2012) found older driver (65+) significantly increased the probability of fatal injury in single vesicle crashes. However, Harb et al. (2008) observed that drivers younger than 35 years old are more likely to have evasive actions which result in more severe injury. Moreover, there are some other studies focused only on younger driver's injury severity. For example, Weiss et al. (2014) analyzed crashes involving younger drivers and identified factors associated with their injury severity. The results show that young drivers' risky behavior, the presence of passengers and the involvement of vulnerable road users are the three main contributors to crash severity in both single-vehicle and two-vehicle crashes.

In general, females are more likely to face fatal injury than males (Kockelman and Kweon, 2001; Habib and Forbes, 2014). But Srinivasan (2002) observed that risk for females was higher than males only for mild injury level, but there was no significant difference between males and females at higher injury severity levels. Moreover, Kim et al. (2012) claimed that there is a higher probability of fatality for male drivers in a newer vehicles compared with females, although newer vehicles can reduce injury severity in single-vehicle crashes. This result indicated that the safety benefit of the newer vehicle is offset by more aggressive driving behavior. Weiss et al. (2014) investigated not only drivers themselves, but also influence of the passengers' gender on younger driver. They found that the presence of passengers - in particular, young male or a group of young males and females - significantly increased the probability of serious and fatal injury. For example, compared with no passengers, companion with only female passengers doubles the driver injury severity for serious injury and triples for fatality. The likelihood of fatal injury increased for more than 5 times when a group of passengers was in a vehicle.

Driver conditions have been found to affect injury severity. Nassiri and Edrissi (2006) found that driver fatigue has the highest negative effect on injury severity in truck crashes for two-lane rural highways in Iran using ordered logit model. For large truck drivers, fatigue may also result in more severe injury (Zhu and Srinivasan, 2011). Williamson et al. (2011) reviewed previous studies on the relationships between major causes of fatigue (sleep homeostasis factors, circadian influences and nature of task effects) and injury severity. Although they found these major causes had adverse effects on driving performance, they could not find sufficient evidence to support a direct link between circadian-related fatigue and injury severity. Moreover, Zajac and Ivan (2003) found that drinking and driving can significantly increase the risk of fatal crashes. Chang and Chien (2013) also found that if a driver was drinking and driving without seatbelt usage, the predicted level of injury was more likely to be fatal. This result showed that drinking and driving without seatbelt increase the risk of fatality. Similarly, Zhu and Srinivasan (2011) reported that driver fatigue, illness, distraction and unfamiliarity with vehicle significantly increase injury severity.

Use of restraint devices is also associated with injury severity. Bedard et al. (2002) found that seatbelts or helmets significantly reduced injury severity. On the other hand, Srinivasan (2002) found that injury severity in a crash in which an air bag was deployed was higher than a crash in which an air bag was not deployed. This is because air bag is usually deployed at high impact speed where drivers are more likely to be severely injured.

Vehicle rollover generally increases injury severity. Khattak et al. (2003) found that rollover leads to more severe injuries in single-truck crashes. They found that dangerous driving behavior (speeding, reckless driving, alcohol or drug habit etc.), left- or right-turning and curved road were associated with higher probability of rollover. Srinivasan (2002) claimed that tripped rollover will result in nearly eight time higher chance of fatal injury for moped riders, compared to non-rollover.

Also, presence of passengers affects driver behavior and driver injury severity. Lee and Abdel-Aty (2008) found that drivers tend to drive safer and less likely to be fatal/severely injured when they are accompanied by passengers and carry more than one passenger. However, Neyens and Boyle (2008) found that teenage drivers distracted by passengers displayed unsafe behavior (e.g. speeding) and they are more likely to be severely injured. Some similar results have also been found by Weiss et al. (2014) for motorcycle drivers.

Some studies investigated influence of vehicle characteristics on injury severity. Harb et al. (2009) found that truck drivers are more likely to perform evasive actions to avoid crashes compared to passenger car drivers. This may be due to the fact that truck drivers benefit from professional driver training programs. Moreover, drivers are more likely to take evasive actions at higher speed limits compared to lower speed limits due to driver's higher alertness on higher speed limit roads (Harb et al., 2009).

In addition, since crash injury severity increases with the mass of vehicles and speed limit at the crash site (Sobhani et al., 2011) and collision force (Wang and Qin, 2014), collisions with trucks increase injury severity than collisions with passenger cars (Duncan et al., 1998). Similarly, Zhu and Srinivasan (2011) found that head-on collisions between truck and car were the most dangerous crash type. Helai et al. (2008) had similar results that heavy vehicles have a better resistance on crashes and thus induce less severe injuries.

Truck body type was also found to affect injury severity for collisions among trucks. For example, Chen and Chen (2011) also found that trucks hauling a trailer with heavy cargo result in more severe injuries compared with light heavy cargo trucks and single-unit trucks. They also found that a single-unit truck has lower probability of severe injury than all other non-single-unit trucks in single vehicle crashes, but it has higher probability of severe injury than the other types of truck in multi-vehicle crashes. Lemp et al. (2012) found similar results for the truck and trailer.

Model year of vehicles is another important factor associated with injury severity because new vehicle technology improves vehicle protection with more advanced materials and equipments. Khorashadi et al. (2005) developed MNL model using a four-year crash data in California and found that 1981 or older model years of cars are more likely to cause severe or fatal injury. Similarly, Rana et al. (2010) found drivers of older vehicles (over ten years) may have higher injury severity level than those of new vehicles, due to the advance in vehicle and safety design. Kim et al. (2012) claimed that newer vehicles can reduce injury severity in single-vehicle crashes, but male drivers are more likely to be severely injured than female driver in new vehicles in single-vehicle crashes.

Some researchers also investigated the effect of vehicle movement. Wang and Abdel-Aty (2008) examined left-turn crash injury severity in central Florida. They found that left-turning traffic colliding with opposing through traffic and with near-side through traffic may result in more severe injury compared with the other left-turn crashes.

The effect of road geometry on injury severity has also been examined. Chung (2013) found that fatality is associated with narrower median islands and the fixed object in the median islands increases injury severity. Moreover, Zhu and Srinivasan (2011) found that crashes on roadways with more number of lanes would result in less severe injury and crashes on roads with higher speed limit would result in more severe injury. Grades also have some influence on injury severity. Lemp et al. (2012) found that grades of 2% uphill and downhill increased injury severity. But in some cases, they decreased injury severity. For example, when a large truck is maneuvering a curve in the road, the probability of fatality is predicted to drop (Lemp et al., 2012). The potential reason is that some complex road geometry conditions increase driver awareness and encourages more cautious driving. Huang et al. (2008) investigated crashes that occurred at intersections. The results show that when drivers on the minor road merge into the major road at three-leg intersections, they have a higher probability of colliding with vehicles on major road and this results in more severe injury. They also found that nighttime, right-most lane, and red light cameras installed at intersections are associated with more severe injury. Recently, Geedipally (2014) studied crashes on ramps and at crossroad ramps terminals. They found that crashes on ramp segments with two lanes tend to be less severe than the crashes on ramps with one lane.

Some environmental factors such as lighting and road surface conditions are also found to be closely related to injury severity. Khorashadi et al. (2005) claimed that crashes in the morning (5:31-8:00) are less likely to result in severe or fatal injury in both urban and rural areas. However, Islam and Hernandez (2014) found that clear sky condition results in greater probability of fatalities (204.5%) but less likelihood of incapacitating injury (48%) in urban areas. This is because drivers tend to drive faster under clear sky condition due to good visibility. Rana et al. (2010) found that driver injury severity was lower when crashes occurred on icy road surface than dry or wet road surface. Zhu and Srinivasan (2011) found that crashes in dark conditions with lighting lead to most severe injury, but the injury is less severe on wet road surface. Similarly, higher probability of more severe injury at nighttime was also found by Weiss et al. (2013) and Helai et al. (2008).

2.2. Single-Vehicle Crashes

Some studies focused on single-vehicle crashes to identify their unique characteristics. These studies commonly found that injury severity in single-vehicle crashes is associated with driver's error or abnormal behavior such as distraction, alcohol/drug use, non-seat-belt use and speeding. Anowar et al. (2012) examined the effects of different factors on the severity of single-vehicle crashes that occurred during holidays in Canada and found that no restraint use, driver violations and errors, alcohol use or fatigue were highly associated with more severe injury. Moreover, Jiang et al. (2013) found out that nighttime was associated with lower probability of severe injury but there was no significant difference in injury severity between nighttime and daytime. This is because drivers tend to drive more carefully at night due to adverse driving condition. However, traffic volume is usually lower at night, and this may encourage drivers to drive at higher speeds. These complexities may cancel out the positive effect of driver's careful attention. Xie et al. (2012) found that automobile drivers usually sustain less severe injury than van in single-vehicle crashes. For crashes in work zones, drivers are more likely to sustain incapacitated and fatal injuries. Kim et al. (2012) reported that seatbelt use reduced the probability of serious injury in crashes but other risky behavior such as drinking and driving, while cell phone use increased the probability of serious injury.

2.3. Two-Vehicle Crashes

Some studies focused on two-vehicle crashes only. For instance, Duncan et al. (1998) investigated injury severity of truck-passenger car rear-end crashes using an ordered probit model. Zhu and Srinivasan (2011) analyzed injury severity of different collision types of car-truck crashes and found that injury severity was higher for head-on and sideswipe crashes. However, these studies did not report injury severity of the other types of two-vehicle crashes (e.g. car-car and truck-truck crashes). More recently, Abay et al. (2013) considered characteristics of both vehicles involved in two-vehicle crashes to estimate injury severity. The study found that lighter vehicle's driver is more likely to be severely injured than heavier vehicle's driver. However, it did not examine the difference in injury severity among different types of two-vehicle crashes. Qin et al. (2013) compared injury severity between car-truck and truck-truck crashes but could not find a significant difference in spite of differential impacts of collisions. Sobhani et al. (2011) combined Newtonian Mechanics and Generalized Linear Regression model to investigate two-vehicle crashes in Australia. The study identified the relationship among crash impact type in terms of collision angle, presence of air bag and/or seat belt, and occupant's age. They found that in general, the presence of air bag and seat belt reduced the crash injury severity. However, in some conditions such as certain collision angles and older driver group, injury severity was higher than expected. Jiang et al. (2013) found that light-truck-involved crashes produced less severe injury than car-car crashes but could not find a significant difference in injury severity between car-car crashes and heavy-truck-involved crashes. Torrão et al. (2014) reported that the engine size of the partner vehicle affects serious injury and fatality in the vehicle. However, the study only considered vehicle characteristics (e.g. age, weight, speed), but not characteristics of occupants in the vehicles.

2.4. Comparison of Single- and Two-Vehicle Crashes

Some studies found differences in injury severity between single-vehicle and two-vehicle/multi-vehicle crashes. For example, Khorashadi et al. (2005) reported that drivers are more likely to be severely injured in multi-vehicle crashes compared with single vehicle (truck) crashes in rural areas. But they did not differentiate truck driver injury severity or passenger vehicle. Wang and Kockelman (2005) found opposite effects of several variables such as curb weight, lighting condition, and grade on driver injury severity between single-vehicle and two-vehicle crashes using Heteroscedastic Ordered Logit model. Savolainen and Mannering (2007) found that helmet use is more likely to lower motorcyclist's fatality for right-angle multi-vehicle crashes, but not fatality of single-vehicle crashes using a nest logit model. Chen and Chen (2011) also found that the effects of old drivers and truck cargo defect on injury severity were opposite between the single-vehicle and multi-vehicle truck-involved crashes using mixed logit model. Weiss et al. (2014) reported that injury severity of young drivers (15-24) in larger vehicles is higher in single-vehicle crashes but lower in two-vehicle crashes compared to young drivers in smaller vehicles using MXL models. However, these studies focused on a specific two-vehicle crashes (e.g. crashes involving motorcycles or trucks only) or did not clearly show the types of vehicles involved in each two-vehicle crash.

2.5. Limitations of Past Studies

Based on this literature review, it was found that there has not been a study that comprehensively evaluates injury severity for two-vehicle crashes considering different types of vehicles involved in crashes. Variations in driver injury severity in two-vehicle crashes are more complex due to the different collision types (head-on, rear-end, etc.) and vehicle body types (car-car crashes, truck-car crashes, truck-truck crashes.).

Identifying the critical factors affecting injury severity and their real effects on injury severity is challenging. Thus more advanced models should be developed for this task. Moreover, the models should be developed separately for single- and two-vehicle crashes so that the difference in the significant factors and their influence on injury severity can be identified. Thus, there is a need for more extensive study on characteristics of injury severity of single-vehicle crashes and different types of two-vehicle crashes with both parametric models and non-parametric models. In addition, the performance of models should be compared.

3. Data

3.1. Description of Data

A five-year (2004-2008) crash record provided by the Ministry of Transportation Ontario was used in this study. This data consist of crash data, traffic volume data and road geometry data for provincial highways in Ontario, Canada. The crash data include information on the time of a crash, drivers/passengers and types of vehicles involved in crashes including injury severity, weather/surface conditions at the time of crash, and collision types. Five levels of injury severity are shown in Table 3-1. The location of each crash was identified as a roadway segment designated in LHRS (Linear Highway Referencing System) number. MTO's LHRS data include road geometric characteristics and average traffic volume of each roadway segment. Table 3-2 summarizes the list of variables included in the data.

Table 3-1. Injury Severity Levels in Ontario (Source: Ministry of Translation Ontario, 2012)

Level of injury severity	Description
Fatal injury	Person was killed immediately or within 30 days of the motor vehicle collision.
Major injury	Person was admitted to hospital.
Minor injury	Person went to hospital and was treated in the emergency room but was not admitted.
Minimal injury	Person did not go to hospital when leaving the scene of the collision, Includes minor abrasions, bruises and complaints of pain.
No injury	No person was injured.

Note: Higher injury level is more severe injury with fatal injury being the highest and no injury being the lowest.

This study analyzes single-vehicle and two-vehicle crashes involving at least one injury (driver or passenger). A total of 13,880 single-vehicle crashes and 15,556 two-vehicle crashes have occurred during the five-year period. Due to low frequency of fatality (1.3% of total single- and two-vehicle crashes), fatal and major injuries were combined into one category of injury severity. Therefore, four injury severity levels were considered in the analysis.

Table 3-2. Description of Variables

Type of variables	Variable	Symbol	Description
Crash characteristics	Season	Month0	Spring (Mar.-May) Summer (Jun.-Aug.) Fall (Sep.-Nov.) Winter (Dec.-Feb.)
	Day of week	Weekends	Weekdays Weekend
	Time of day	Daytime	Daytime (6:00-18:00) Nighttime (19:00-6:00)
Driver characteristics	Driver action	Dr_Act	Proper action Improper action (e.g. speed too fast, following too close)
	Driver condition	Driver_Con0	Normal condition Abnormal condition (e.g. alcohol or drug use, fatigue)
	Driver age	Driver_Age	Young (≤ 30) Middle1 (31-45) Middle2 (46-60) Old (61 and over)
	Driver sex	Driver_Sex	Female Male
	Injury	Injury_Sev	Fatal and major injury Minor injury Minimal injury No injury
	Safety equipment	Safe_Equ	Not used Used
	Ejection	Ejection	Not ejected from vehicle Ejected from vehicle
Environmental characteristics	Lighting	Lighting0	Good lighting Dark lighting Other lighting conditions
	Weather	Climat	Clear weather Not clear weather
	Road surface condition	Road_Surface0	Dry road surface Wet surface All other road surface conditions

Table 3-2. Description of Variables (Continued)

Type of variables	Variable	Symbol	Description
Vehicle characteristics	Vehicle type	Vehicle_Type0	Passenger car Truck Others
	Model year	Model_Year0	2004-2009 1999-2003 1993-1998 Prior to 1992
	Vehicle age	Veh_age	Vehicle age in the year of crash
	Vehicle movement	Vehicle_Movement0	Going ahead Other vehicle movement
Road geometric characteristics	Speed limit	POSTED_SPEED20	Posted speed limit (km/h)
	Number of lanes	NUM_LANES20	Number of lanes on the road
	Number of Streams	STREAMS	Single stream (undivided highway) Two streams (divided highway) Four streams (Core/collector system)
	Median	MEDIAN0	Grass Other
	Shoulder	Shoulder0	Paved Partly paved Gravel
	Road Surface	SURFACE0	Concrete Gravel/Sand Bituminous and Other
	Median shoulder width	MED_SHLDWIDTH0	Width of the left side shoulder (m)
	Median width	MEDIAN_WIDTH0	Width of median on the road (m)
	Shoulder width	SHLD_WIDTH0	Width of the right side shoulder (m)
	Surface width	SURF_WIDTH0	Width of drivable surface excluding medians or shoulders (m)
	Terrain	TERRAIN	Flat terrain Rolling terrain
	Traffic control	SIGNALS	No signal With signal Others
	Impact	Impact0	Single vehicle impact Other impacts
	Alignment	Alignment0	Curve road Straight road
	Road character	Rd_Char0	Divided road Undivided road
	Functional class	FUNC_CLSS	Freeway, Arterial Collector, Local
	Road type	Road_Type0	Asphalt, All other type
Traffic characteristics	Traffic volume	AADT0	Annual average daily traffic (AADT) for roadway segment (vehicles/day)
	Truck percentage	Truc	Truck percentage in AADT (percentage)

Since there are many people involved in each crash, only driver record was selected in each crash. This is because the impact of collision varies across person's different positions in the vehicle. For instance, when a head-on collision occurs, occupants (driver or passenger) in the front seats are more likely to be severely injured than those in the rear seats. By selecting driver records only, the effect of person's position in the vehicle will be eliminated. The vehicles involved in crashes are classified into four categories: passenger car, light truck, heavy truck and others. Light truck includes passenger van, buses, school vehicle, fire vehicle and pickup truck. Heavy truck includes tractor-trailer, tow trucks, and farm tractor. A majority of heavy trucks (70%) are tractor-trailers. Others include motorbike and motorcycle. A majority of other vehicle type (99%) are motorcycles.

For two-vehicle crashes, each crash has unique geometric, weather and traffic characteristics, which are common to both drivers involved in the crash. Thus, if both driver records are used, these characteristics will be duplicated in the data. To avoid this duplication, the driver record for only one of two vehicles was randomly selected. Since driver's citation record was not provided, the driver at fault was unknown. Injury severity is expected to be different among passenger car (C), light truck (L) and heavy truck (H) drivers involved in the same crash due to difference in weights of vehicles and impact of collisions on vehicle bodies. Thus, the driver record was separated into nine data sets as follows:

1) Car-Car crashes for Car drivers record (C-C)
2) Car-Heavy truck crashes for Car drivers record (C-H(C))
3) Car-Heavy truck crashes for Heavy truck drivers record (C-H(H))
4) Heavy truck-Heavy truck crashes for Heavy truck drivers record (H-H)

5) Car-Light truck crashes for Car drivers record (C-L(C))
6) Car-Light truck crashes for Light truck drivers record (C-L(L))
7) Light truck-Light truck crashes for Light truck drivers record (L-L)
8) Light truck-Heavy truck crashes for Light truck drivers record (L-H(L))
9) Light truck-Heavy truck crashes for Heavy truck drivers record (L-H(H))

The proportions of four injury severity levels in total number of crashes were compared among the 9 two-vehicle crash types. It was observed that when different vehicle categories (i.e. different vehicle size and weight) are involved in crashes, the proportions of fatal/major crashes were generally higher for the drivers of smaller and lighter vehicles. Also, when two vehicles in the same categories (i.e. similar vehicle size and weight) were involved in crashes, injury severity was higher for collisions between larger and heavier vehicles. The structure of the data is shown in Figure 3-1.

3.2. Preliminary Analysis of Data

The association of explanatory variables with injury severity was investigated using two-way chi-square tests. Table 3-3 shows the relationship between the variables and driver's injury severity for single-vehicle crashes. The association is statistically significant at a 95% confidence interval ($p \leq 0.05$). For instance, it was found that fatal and major injuries are more likely to occur at nighttime than daytime (Table 3-3(a)). This is because drivers are more likely to make errors due to shorter sight distance and they tend to drive faster due to lower traffic volume at night. Similarly, drivers are more likely to be severely injured in clear weather conditions when they feel more comfortable with traveling at higher speed (Table 3-3(b)).

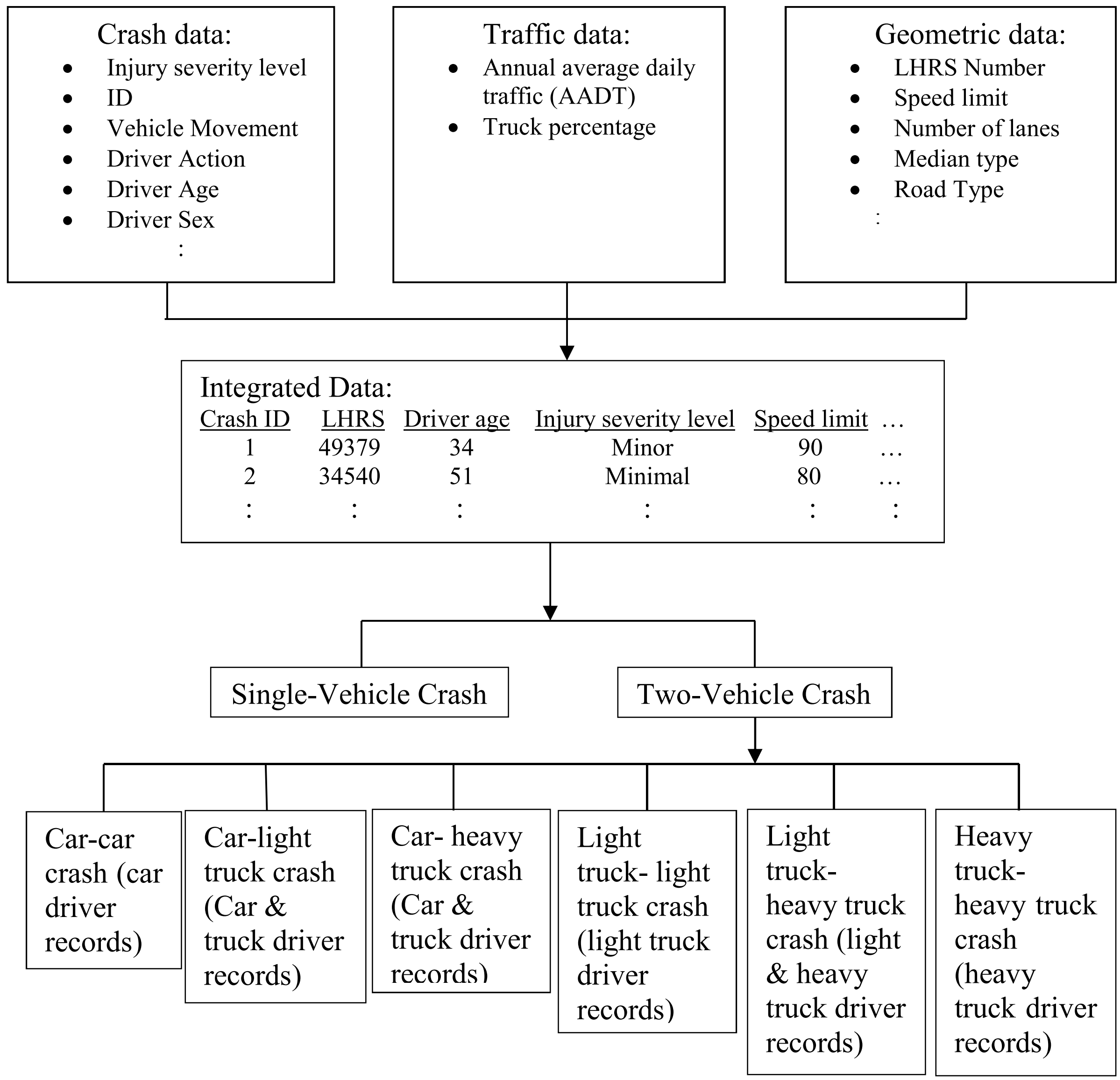


Figure 3-1. Structure of Data

Table 3-3. Relationship between Injury Severity and Explanatory Variables (Single-Vehicle Crashes)

Variable	Injury Severity			
	No injury	Minimal injury	Minor injury	Fatal/major injury
(a) Time of day				
Daytime	795 (10) [a]	3439 (41)	3452 (42)	615 (7)
Nighttime	555 (11)	2013 (40)	1983 (40)	425 (9)
(b) Weather				
Clear	823 (9)	3774 (43)	3350 (38)	838 (10)
Not clear	527 (12)	1678 (37)	2085 (46)	202 (5)
(c) Speed limit				
< 75 km/h	116 (16)	286 (39)	275 (38)	52 (7)
75-80 km/h	294 (9)	1399 (42)	1364 (41)	299 (9)
80-90 km/h	216 (10)	838 (39)	904 (42)	192 (9)
90-100 km/h	724 (10)	2929 (42)	2892 (41)	497 (7)
(d) Number of lanes				
≤ 3	608 (9)	2644 (41)	2661 (41)	567 (9)
4-5	368 (10)	1615 (42)	1505 (40)	318 (8)
6-8	244 (12)	812 (41)	825 (42)	103 (5)
9 and above	130 (13)	381 (38)	444 (44)	52 (5)
(e) Driver age				
Young (≤ 30)	544 (10)	2073 (39)	2340 (44)	368 (7)
Middle1 (31-45)	360 (9)	1642 (42)	1618 (41)	299 (8)
Middle2 (46-60)	297 (11)	1176 (43)	1020 (37)	237 (9)
Old (61 and over)	149 (11)	561 (43)	457 (35)	136 (10)
(f) Driver sex				
Female	371 (7)	2252 (44)	2241 (44)	288 (6)
Male	979 (12)	3200 (39)	3194 (39)	752 (9)
(g) Safe equipment				
Not used	12 (3)	134 (37)	54 (15)	158 (44)
Used	1338 (10)	5318 (41)	5381 (2)	882 (7)
(h) Ejection				
Eject	4 (1)	246 (42)	79 (13)	262 (44)
No eject	1346 (11)	5206 (41)	5356 (42)	778 (6)
(i) Driver action				
Improper	1036 (10)	4567 (42)	4368 (40)	933 (9)
Proper	314 (13)	885 (37)	1067 (45)	107 (5)
(j) Driver condition				
Abnormal	342 (7)	2106 (45)	1668 (35)	596 (13)
Normal	1008 (2)	3346 (39)	3767 (44)	444 (5)

[a]The numbers in parentheses are the proportions of each injury severity level for given category of each explanatory variable.

It was also found that injury severity was lower at the lowest and highest categories of posted speed limits (< 75 km/h and 90-100 km/h, respectively) (Table 3-3(c)). This indicates that drivers tend to drive more cautiously on the highways with higher speed limit above 90 km/h.

It was also found that injury severity was higher at the locations with lower number of lanes (Table 3-3(d)). This indicates that drivers have lower chance of avoiding collision and severe injuries when there is less space available on roadways. Driver demographic characteristics and conditions were also significantly related to injury severity. The result shows that fatal and major injuries are more likely to occur if drivers are older and male, they do not wear safety equipment, they are ejected from the vehicle, and their driving actions and conditions are not normal (Tables 3-3(e)-(j)).

The proportions of four injury severity levels in total number of crashes were also compared among the 9 two-vehicle crash types as shown in Table 3-4. It was observed that when different vehicle categories (i.e. different vehicle size and weight) were involved in crashes, the proportions of fatal/major crashes were generally higher for the drivers of smaller and lighter vehicles. Also, when two vehicles in the same categories (i.e. similar vehicle size and weight) were involved in crashes, injury severity was higher for collisions between larger and heavier vehicles.

Table 3-4. Proportions of Four Injury Severity Levels in Two-vehicle Crashes

Crash type		No injury	Minimal injury	Minor injury	Fatal/major injury
C-C	Frequency	3453	924	1122	181
	%	60.8	16.3	19.8	3.2
C-H(C)	Frequency	214	681	716	191
	%	11.9	37.8	39.7	10.6
C-H(H)	Frequency	1625	70	100	8
	%	90.1	3.9	5.6	0.4
H-H	Frequency	66	35	43	29
	%	38.2	20.2	24.9	16.8
C-L(C)	Frequency	1098	847	1049	196
	%	34.4	26.6	32.9	6.1
C-L(L)	Frequency	1810	560	724	99
	%	56.7	17.5	22.7	3.1
L-L	Frequency	366	116	128	40
	%	56.3	17.9	19.7	6.2
L-H(L)	Frequency	54	132	128	66
	%	14.2	34.7	33.7	17.4
L-H(H)	Frequency	318	26	31	5
	%	83.7	6.8	8.2	1.3

4. Methods

In this chapter, both parametric and non-parametric models for identifying the factors contributing to injury severity and estimating their effects are described. Theoretical backgrounds of each model are explained below.

4.1. Parametric Models

4.1.1. Ordered Logit Model

To account for ordinal nature of injury severity levels (i.e. higher level indicates more severe injury), ordered choice models were utilized. Ordered choice models describe injury severity level as a response variable in a function of explanatory variables. The injury severity level is determined by the following latent measure (Aitchison and Silvey, 1957):

$$U_i^* = \beta X_i + \varepsilon_i \tag{4-1}$$

where

U_i^* = latent and continuous measure of injury severity for driver i;

β = a vector of coefficient for explanatory variables;

X_i = a vector of explanatory variables associated with driver i and crash;

ε_i = random error term.

In the above equations, the random error term reflects unobserved effects of other variables not included in the model on injury severity. If the error term follows a Gumbel distribution, the model is called the ordered logit (OL) model. If the error term follows a normal distribution, the model is called the ordered probit (OP) model.

From the observed injury severity levels {1(no injury), 2(minimal injury), 3(minor injury), 4(fatal/major injury)} in crash records, U_i^* is determined as follows:

$$U_i^* = \begin{cases} 1 & \text{if } U_i^* \leq 0 \text{ (no injury)} \\ 2 & \text{if } 0 \leq U_i^* \leq \mu_1 \text{ (minimal injury)} \\ 3 & \text{if } \mu_1 \leq U_i^* \leq \mu_2 \text{ (minor injury)} \\ 4 & \text{if } \mu_2 \leq U_i^* \leq \infty \text{ (fatal/major injury)} \end{cases} \tag{4-2}$$

where μ's are threshold parameters. The probability $P_i(N)$ that driver i's injury severity is equal to N = 1, 2, 3, or 4, can be calculated as follows:

$$P_i(1) = P_r(U_i = 1) = P_r(U_i^* \leq \mu_1) = P_r(\beta X_i + \varepsilon_n \leq \mu_1) = P_r(\varepsilon_n \leq \mu_1 - \beta X_i)$$
$$P_i(2) = P_r(U_i = 2) = P_r(\mu_1 \leq U_i^* \leq \mu_2) = P_r(\varepsilon_n \leq \mu_2 - \beta X_i) - P_r(\varepsilon_n \leq \mu_1 - \beta X_i)$$
$$P_i(3) = P_r(U_i = 3) = P_r(\mu_2 \leq U_i^* \leq \mu_3) = P_r(\varepsilon_n \leq \mu_3 - \beta X_i) - P_r(\varepsilon_n \leq \mu_2 - \beta X_i)$$
$$P_i(4) = P_r(U_i = 4) = P_r(\mu_3 \leq U_i^* \leq \mu_4) = P_r(\varepsilon_n \leq \mu_4 - \beta X_i) - P_r(\varepsilon_n \leq \mu_3 - \beta X_i)$$

In general, the probability $P_i(N)$ can be calculated using the following equation:

$$\mathrm{P_i(N)} = \phi(\mu_n - \beta X_i) - \phi(\mu_{n-1} - \beta X_i) \tag{4-3}$$

where

$\phi(.)$ = cumulative distribution function of the logistic distribution.

The parameter β shows the effect of explanatory variables on injury severity. A positive sign of β indicates higher injury severity as the value of the associated variable X increases and vice versa. The coefficients are estimated by using the method of maximum likelihood. A measure of goodness-of-fit is as follows:

$$\rho^2 = 1 - \left[\frac{\ln L_b}{\ln L_0}\right] \tag{4-4}$$

where

$\ln L_b$ = the log likelihood at convergence;

$\ln L_0$ = the log likelihood computed at zero.

The ρ^2 value varies between zero and one, and higher value closer to one indicates a better model fit.

4.1.2. Heteroscedastic Ordered Logit Model

In the conventional ordered logit model, the variance in the error term is assumed to be the same for all observations (i.e. crashes and drivers) or homoscedastic. However, this assumption is violated if the unobserved effects of variables (i.e. error terms) on driver's injury severity are different for different crashes and drivers.

Unlike the OL model, the heteroscedastic ordered logit model (HOL) allows the error term to vary for each observation as follows (Wang and Kockelman, 2005):

$$P_i(N) = \phi\left(\frac{\mu_n - \beta X_i}{\sigma_i}\right) - \phi\left(\frac{\mu_{n-1} - \beta X_i}{\sigma_i}\right) \quad (4\text{-}5)$$

where

${\sigma_i}^2$ = the variance of driver i's random error term (ε_i).

This variance is described in a function of the variables associated with the variance of driver i's error term, Z_i, as follows:

$${\sigma_i}^2 = [\exp(\gamma Z_i)]^2 \quad (4\text{-}6)$$

where

γ = coefficients for variable Z_i.

In the conventional ordered models, the coefficient γ is set to zero. The coefficients β and γ are estimated using the maximum likelihood method. Higher variance indicates higher uncertainty with driver's injury severity for a given value of the variable Z_i (Lemp et al., 2012). Thus, the HOL model can better reflect the variance in unobserved effects of a variable on driver's injury severity across observations or heteroscedasticity. The HOL models were estimated using SAS 9.2 (SAS Institute, 2012). In SAS, the variable(s) with heteroscedasticity can be specified separately using the procedure known as "qlim" .

4.2. Non-parametric models

4.2.1. Classification and Regression Tree

The classification and regression tree (CART) is a data mining technique to find complex relationships among different variables. Unlike parametric models, there is no pre-defined relationship between the dependent variable and independent variables. The model does not require pre-processing of the data (e.g. dummy variables) to identify the association of independent variables with a dependent variable. An advantage of CART is its ability to avoid multi-collinearity problems and isolate outliers (Chang and Wang, 2006). Unlike other data mining methods such as neural networks, tree structures make interpretation of the results easier (Pande et al., 2010).

In the CART, tree structures are developed in the following process. The tree growing process creates groups by partitioning samples such that samples within the same group are as homogenous (pure) as possible. For this purpose, several split rules can be applied to generate nodes and branches in the tree structures. These splits are evaluated and ranked based on the Gini reduction criterion, which measure the "worth" of each split to achieve maximum homogeneity (Pande et al., 2010). The worth in the Gini reduction criterion or the Gini measures is determined based on the "impurity" of each node which reflects the degree of non-homogeneity of samples in each node. As the samples in the same node are more homogeneous, the Gini measures decrease. Thus the objective of the splits is to minimize the Gini measures or maximize the homogeneity. This tree growing process stops when the number of observations in a node is equal to a pre-specified minimum or the reduction in the Gini measures is less than a pre-specified minimum.

Some studies also predicted injury severity using the CART. Chang and Wang (2006) predicted injury severity of crashes in Taiwan using the CART and found that vehicle type was strongly related to crash injury severity. Montella et al. (2012) found from the result of the CART that road type was significantly associated with injury severity of powered two-wheeler crashes in Italy. Eustace et al. (2014) also applied the CART to prediction of injury severity of run-off-road crashes in Ohio and found that road condition was the most important factor.

Developing a tree using single data set may cause overfitting problem, which makes it difficult to classify different data sets using the tree (Chang and Wang, 2006). A remedy is to use for 70% of the data training and constructing a tree while leaving 30% of the data for validation. The CART was developed using the SAS Enterprise Miner 6.2 (SAS Institute, 2009).

4.2.2. Random Forests

In this study, the random forests method is used to determine the ranking of importance of variables in the prediction of driver's injury severity and identify inputs of independent variables before developing the CART. The random forests method determines the ranking using unpruned classification or regression trees created by randomly selecting samples with replacement (i.e. bootstrapping) (Ho, 1995). The procedure of determining variable importance in the random forests method is described as follows:

1. Select a bootstrap sample.
2. Grow a classification tree to fit to the bootstrap sample so that the variable can be selected only from a small subset of randomly selected variables for each split in the classification tree.
3. Predict the response variable for the samples not selected in the bootstrap sample (i.e. out-of-bag samples) using the classification tree in Step 2. The response variable is predicted as the classification category of the variable with the highest proportion of samples.
4. Compare the observed and predicted categories of the response variable to calculate the misclassification rate (accuracy) of the tree.
5. For each predictor variable, permute the value of the variable in the out-of-bag samples. Predict the response variable using the classification tree in Step 2 to calculate the new misclassification rate of the tree. The importance score for each variable is computed based on the difference between the misclassification rates before and after the permutation (Strobl et al., 2007). For instance, higher

difference between the two misclassification rates increases the importance score – i.e. variable importance is higher.

6. Repeat Steps 1-5 until a sufficiently large number of trees are grown using different bootstrap samples. Calculate the average importance score for each variable in different trees.

The random forests method was applied using the R software (R Development Core Team, 2006)

4.2.3. Boosted Regression Trees

The boosted regression trees model (BRT) is a tree-based model which improves the performance of a single tree model (CART). The BRT can handle different types of predictor variables and accommodate missing data (Elith et al., 2008). The BRT does not require prior data transformation or elimination of outliers similar to the CART. The BRT can also fit complex nonlinear relationships, and automatically handle interaction effects between predictors.

A main disadvantage of the CART is that the tree structure significantly changes even if there is a small change in data (Chung, 2013). Thus, the CART is unstable when handling crash injury severity data with high variance. Although increasing the complexity of tree structures by adding more split variables or increasing depth of trees will decrease bias in predictions, it will also increase variance in predictions (De'ath, 2007). Thus, the "bagging" technique is used for more complex trees with higher variance and lower bias. This technique includes the following steps:

1. Take a bootstrap sample from the data set.
2. Fit the tree to this data set.
3. Repeat step 1 and 2 (typically 50-1,000 times).
4. Predict for new data using each of the fitted models and average the predictions.

Similar to the bagging technique, the BRT balances the bias and variance in predictions. However, unlike the bagging technique, the BRT sequentially applies a higher weight to incorrectly classified observations and a lower weight to correctly classified observations as a series of trees are fit to bootstrap samples. In this "boosting" process, the weights of the observations that are more difficult to be classified will increase. Thus, the BRT will increase the chance that the observations with higher weight are correctly classified (De'ath, 2007).

In the BRT, a basis function $f(x)$ which describes a response variable y in a function of explanatory variables x is expressed as a sum of the basis functions for individual trees as follows (Hastie et al., 2009):

$$f(x) = \sum_m \beta_m b(x;\, \gamma_m) \qquad (4\text{-}7)$$

where $b(x;\, \gamma_m)$ is a basis function for individual tree m, γ_m is the split variables, their values at each node and the predicted values, and β_m is the parameter estimated such that the squared error $(y - f(x))^2$ is minimized (De'ath, 2007). The squared error is one type of the loss function, $L(y, f(x))$.

To estimate the sum of the basis functions in Equation (4-7), Friedman (2001) developed gradient boosting. In gradient boosting, an initial basis function is set to zero. Then this basis function is updated as a series of trees are fit as follows:

1. For a least-squares regression tree m, estimate γ_m of the basis function $f_m(x)$ and calculate the residuals (i.e. the derivative of a loss function).

2. Estimate β_m such that it minimizes the following overall loss:

   $L(y, f_{m\text{-}1}(x) + \beta_m b(x; \gamma_m))$

   where $f_{m\text{-}1}(x)$ is the basis function for the previous tree m-1. In this procedure, gradient boosting adjusts the weight (β) of the current tree based on the prediction in the previous tree.

3. Calculate overall basis function $f(x)$ as the sum of $f_m(x)$ as shown in Equation. (4-7).

The BRT has been applied to various study areas including animal ecology (Elith et al., 2008), air pollution (Carslaw and Taylor, 2009), and epidemiology (Cheong et al., In press). Recently Chung (2013) applied the BRT to prediction of injury severity of single-vehicle motorcycle crashes in Taiwan. In particular, the BRT performs better for the injury severity data with relatively smaller sample size of fatal and severe injury crashes than non-severe injury crashes (Chung, 2013). The study found that the BRT showed higher classification accuracy than the CART. However, since the study focused on single-vehicle crashes with a single vehicle type (motorcycle) only, the capability of the BRT for predicting injury severity for single-vehicle crashes with different vehicle types and multi-vehicle crashes is still unknown. In this study, the BRT was developed using the R software with the *dismo* package (Elith and Leathwick, 2014).

5. Results and Discussion

5.1. Heteroscedastic Ordered Logit Model

Heteroscedastic ordered logit (HOL) models were developed for single-vehicle and two-vehicle crashes separately. Tables 5.1~5.5 show significant variables associated with injury severity and their coefficients for the single- and two-vehicle crash models.

5.1.1. Single-Vehicle Crash Model

The model result for single-vehicle crashes shows that all variables except the variance of random effects for young drivers are significant at a 95% significance level as shown in Table 5-1. The table shows that the injury severity was higher on the roads with higher posted speed limit. In general, higher speed limit implies higher actual speed of vehicle when the crash occurred. Those drivers in these vehicles with higher speed are more likely to experience higher impact from collision, which may result in severe injury. However, drivers are less likely to be severely injured on the road with more lanes. This is because in general, the roads with higher number of lanes usually have better safety facilities (e.g. well-paved surface, better lighting conditions) than those with only one or two lanes. Moreover, drivers can avoid severe collisions more easily when more space (more lanes) is available in the roadway with higher number of lanes. These results are consistent with Zhu and Srinivasan (2011). However, drivers are more likely to be severely injured on curved roads than straight roads due to higher likelihood of losing control and hitting fixed objects on the roadside. Similar effect of curved roads was also reported in Wang and Kockelman (2005).

Drivers in passenger cars, light trucks and heavy trucks are more likely to be severely injured compared to motorcycle riders. This result contradicts the past studies that motorcycle drivers are more likely to be severely injured than passenger car and truck drivers (e.g. Savolainen and Mannering, 2007). It was observed that motorcycle riders have higher proportion of fatal/major injury but lower proportion of minor injury (the next highest injury severity level) than car and truck drivers. Thus, there was no consistent trend of more severe injury for motorcycle riders. It was also observed that heavy truck drivers sustain higher injury severity than car and light truck drivers.

The result of the model also shows that young drivers (≤ 30) are more likely to be severely injured compared to older drivers (> 30) in single-vehicle crashes. This is consistent with the finding of the past studies (e.g. Chang and Yeh, 2007). Thus, they are more likely to make errors and be involved in severe single-vehicle crashes. Compared to male drivers, female drivers are more likely to be severely injured. This is consistent with Wang and Kockelman (2005). Safety equipments reduce the injury severity similar to Chung (2013) whereas ejection from vehicles increases injury severity.

It was observed that the variance of random effects for safety equipments and ejection was significant at a 95% significance level. This indicates that the variance of random effects must be considered in the model. HOL models also provided better model fit than OL models based on higher log likelihood ratio index. The result of variance indicates that there is the largest variance in injury severity for ejection. This is potentially because injury severity can greatly vary depending on whether drivers are fully or partially ejected and whether ejected drivers hit the fixed objects (e.g. tree, median barrier) or not. Similarly injury severity significantly varies with safety equipments as the levels of human body protection are different for various types of equipments (e.g. seat belt, helmet, air bag).

Unexpectedly, none of environmental factors was significant in single-vehicle crashes. This is potentially because environmental factors have mixed effects on driver behaviour. For instance, poor visibility and slippery road in adverse weather can increase chance of driver's judgment errors. On the other hand, these can also increase driver's awareness of driving condition and result in more cautious driving. Thus, the former increases driver's injury severity whereas the latter decreases driver's injury severity.

Table 5-1. Parameters of HOL Model for Single-Vehicle Crashes

Parameter	Estimate	Pr > t
Intercept	1.51	<.0001
Speed limit (km/h)	0.003	0.02
Passenger car (1 = passenger car; 0 = otherwise)	0.32	0.001
Light truck (1 = light truck; 0 = otherwise)	0.29	0.003
Heavy truck (1 = heavy truck; 0 = otherwise)	0.43	<.0001
Young (1 = age ≤ 30; 0 = otherwise)	0.05	0.02
Female (1 = female; 0 = male)	0.06	0.01
Safety equipments (1 = with safety equip.; 0 = no safety equip.)	-0.64	<.0001
Ejection (1 = ejected from vehicle; 0 = not ejected from vehicle)	1.14	<.0001
Number of lanes	-0.02	<.0001
Curved road (1=curved; 0= straight)	0.08	0.004
Variance		
Young	-0.07	0.06
Safety Equipment	-0.78	<.0001
Ejection	0.89	<.0001
μ_1	1.52	<.0001
μ_2	3.27	<.0001
Log likelihood at convergence ($L^*(\beta)$)	-15146	
Log likelihood ratio index (ρ^2)	0.02	
Number of observations	13277	

5.1.2. Two-Vehicle crash models

In the two-vehicle crash models, the effects of variables on injury severity are generally similar to the effects in the single-vehicle crash model as shown in Tables 5-2~5-4.

Female drivers and no use of safety equipments increase injury severity. Due to rare occurrence of driver's ejection from vehicles in two-vehicle crashes, ejection was not included in the models.

In Table 5-2, it is interesting to note that young drivers are less likely to be severely injured in two-vehicle C-C crashes than old drivers (> 61) unlike single-vehicle crashes. These opposite effects reflect that compared to old drivers, young drivers are more likely to take evasive actions to avoid crashes with another vehicle in high traffic volume conditions where two-vehicle crashes occur more frequently. The result also indicates that old drivers are more susceptible to injury than younger driver groups (≤ 60) when their vehicles collide with another vehicle. The opposite effects of older drivers ((≤ 50) on injury severity between single-vehicle and multi-vehicle crashes were also reported in Chen and Chen (2011).

In the C-C crash model, abnormal driver conditions (e.g. alcohol use, fatigue) increase injury severity. However, proper driving actions also increase injury severity of two-vehicle crashes. This reflects that when the driver with proper driving actions is hit by the driver with improper driving actions, he/she cannot usually anticipate the crash occurrence and cannot take evasive actions to avoid crashes. Consequently, this results in the driver's higher injury severity. Injury severity is also higher for crashes at nighttime than daytime. This reflects that drivers make more errors in poor lighting conditions and they tend to drive faster when traffic volume is low at nighttime. Injury severity is lower for newer vehicles as they have better safety equipments and design features which protect drivers.

Table 5-2. Comparison of Heavy-truck-involved Two-vehicle Crash Models with Car-Car Crash Model

	C-C		C-H(C)		C-H(H)		H-H	
Parameter	Estimate	Pr > t	Estimate	Pr > t	Estimate	Pr > t	Estimate	Pr > t
Intercept	2.33	<.0001	4.52	<.0001	-0.91	0.08	2.15	0.01
Female	0.71	<.0001	-[a]	-	-	-	-	-
Young (≤ 30)	**-0.24**	0.009	-	-	-	-	-	-
Middle1 (31-45)	-0.23	0.02	-	-	-	-	-	-
Middle2 (46-60)	-0.20	0.05	-	-	-	-	-	-
Daytime	-0.29	<.0001	-	-	-	-	-	-
Safety Equip.	-1.92	<.0001	-2.58	<.0001	-1.19	0.02	-1.89	0.02
Abnormal cond.	0.16	0.006	-	-	-	-	0.73	0.01
Improper action	-0.61	<.0001	-	-	-0.81	<.0001	-	-
Vehicle age	0.01	0.05	-	-	-	-	-	-
Angle	0.49	<.0001	0.87	0.006	0.81	0.01	-	-
Head-on	1.62	<.0001	3.89	<.0001	1.47	<.0001	-	-
Sideswipe	**0.53**	<.0001	-	-	-	-	**-0.87**	0.02
Asphalt over concrete	-0.28	<.0001	-	-	-	-	-	-
Wet surface	0.13	0.07	-	-	-	-	-	-
Median width (m)	-0.01	0.01	-	-	-	-	-	-
Surface width (m)	-0.01	<.0001	-	-	-	-	-	-
Variance								
Female	0.24	0.004	-	-	-	-	-	-
Head-on	0.69	<.0001	1.66	<.0001	-0.82	0.065	-	-
Angle	-	-	0.80	0.001	-	-	-	-
Abnormal cond.	-	-	-	-	-	-	-	-
μ_1	0.83	<.0001	2.05	<.0001	0.52	<.0001	0.90	<.0001
μ_2	3.22	<.0001	4.72	<.0001	3.14	<.0001	2.34	<.0001
$L^*(\beta)$	-5318		-2025		-672		-216	
ρ^2	0.07		0.06		0.05		0.04	
No. of obs.	5532		1757		1757		170	

[a] A variable is excluded due to statistically insignificance of the variable at a 90% significance level.

The type of collisions between two vehicles is significantly related to injury severity of two-vehicle crashes. Angle, head-on and sideswipe collisions produce more severe injury than the other types of collisions (e.g. rear-end, turning) for most two-vehicle crashes. It should be noted that the effects of sideswipe collisions are different between C-C crashes and H-H crashes as shown in Table 5-2. The result shows that sideswipe collisions between cars result in higher injury severity but sideswipe collisions between heavy trucks result in lower injury severity compared to the other types of collisions. These opposite effects are potentially because when sideswipe crashes occur between heavy trucks, long trailers are more likely to collide each other from the sides whereas drivers in tractors are less influenced by the impact of the collision.

Some geometric and environmental factors were also significant for C-C crashes. Injury severity was higher on the road with narrower median and travel lanes, but lower on the asphalt over concrete pavement. Injury severity was also higher in wet surface conditions than the other surface conditions.

It was observed that the variance of random effects for head-on collisions was significant at a 95% significance level for the C-C and C-H(C) models. This implies that injury severity of drivers involved in head-on collisions significantly vary among different two-vehicle crashes. Thus, their injury largely depends on the variance in the unobserved effects (e.g. point of impact). The result also shows a significant variance of random effects for angle collisions when car drivers are involved in C-H crashes. This indicates that larger differences in size and weight between two vehicles contribute to greater variation in injury severity of the driver in a smaller and lighter vehicle.

C-C crashes were also compared with light-truck-involved two-vehicle crashes as shown in Table 5-3. Similar to heavy-truck-involved crashes, the effects of nighttime, safety equipments, improper action, and head-on collisions are significant. It was observed in C-L crash models that car driver's injury severity is higher for angle and sideswipe crashes, not only head-on crashes, than the other crash types unlike light truck drivers. This indicates that car drivers are more vulnerable than light truck drivers in various crash types. The effect of sideswipe collisions was also negative for L-L crashes similar to H-H crashes due to stronger resistance to impacts of collisions with more rigid vehicle body.

Table 5-3. Comparison of Light-truck-involved Two-vehicle Crash Models with Car-Car Crash Model

Parameter	C-C Estimate	Pr > t	C-L(C) Estimate	Pr > t	C-L(L) Estimate	Pr > t	L-L Estimate	Pr > t
Intercept	2.33	<.0001	3.67	<.0001	1.41	0.001	2.93	<.0001
Female	0.71	<.0001	0.57	<.0001	0.85	<.0001	0.38	0.04
Young	-0.24	0.009	-[a]	-	-	-	-0.58	0.003
Middle1	-0.23	0.02	-	-	-	-	-0.76	<.0001
Middle2	-0.20	0.05	-	-	-	-	-	-
Daytime	-0.29	<.0001	-	-	-	-	-	-
Safety Equipments	-1.92	<.0001	-2.99	<.0001	-1.64	<.0001	-2.67	<.0001
Abnormal condition	0.16	0.006	0.22	0.03	-	-	-	-
Improper action	-0.61	<.0001	-0.97	<.0001	-1.30	<.0001	-	-
Vehicle age	0.01	0.05	-	-	0.03	<.0001	-	-
Angle	0.49	<.0001	0.43	0.0006	-	-	-	-
Head-on	1.62	<.0001	1.61	<.0001	0.95	<.0001	1.08	<.0001
Sideswipe	0.53	<.0001	0.24	0.03	-	-	-1.34	<.0001
Asphalt over concrete	-0.28	<.0001	-	-	-	-	-	-
Wet surface	0.13	0.07	-	-	-	-	-	-
Median width (m)	-0.01	0.01	-	-	-	-	-	-
Surface width (m)	-0.01	<.0001	-	-	-	-	-	-
Weekend	-	-	-0.25	0.004	-	-	-	-
Undivided	-	-	0.40	<.0001	-	-	-	-
Variance								
Female	-0.24	0.004	-	-	0.28	0.01	-	-
Head-on	0.69	<.0001	1.16	<.0001	0.69	<.0001	-	-
Vehicle age	-	-	-	-	-0.02	0.04	-	-
Asphalt over concrete	-	-	-	-	-	-	-	-
Improper action	-	-	-	-	0.44	0.0003	-	-
Sideswipe	-	-	0.33	0.007	-	-	-	-
Abnormal condition	-	-	0.26	0.02	-	-	-	-
μ_1	0.83	<.0001	1.29	<.0001	0.88	<.0001	0.92	<.0001
μ_2	3.22	<.0001	4.16	<.0001	3.45	<.0001	2.78	<.0001
$L^*(\beta)$	-5318		-3694		-3113		-657	
ρ^2	0.07		0.06		0.07		0.09	
No. of obs.	5532		3132		3128		639	

[a] A variable is excluded due to statistically insignificance of the variable at a 90% significance level.

The result also shows that the variance of random effects for head-on collisions was significant in C-L crash models at a 95% significance level. However, there were more variables with heteroscedasticity for C-L crashes than C-H crashes. This indicates that uncertainty with driver's injury severity increases with smaller difference in size and weight between two different types of vehicles.

Two L-H crash models were compared between light truck and heavy truck drivers as shown in Table 5-4. The result shows that angle crashes significantly increase light truck driver's injury severity but not heavy truck driver's. Since angle crashes also significantly increase car driver's injury severity but not light truck driver's as shown in Table 5-4, angle crashes tend to increase injury severity of drivers in smaller and lighter vehicles only.

Table 5-4. Comparison of L-H Crashes between Light Truck and Heavy Truck Drivers

Parameter	L-H(L) Estimate	Pr > t	L-H(H) Estimate	Pr > t
Intercept	4.39	<.0001	0.54	0.53
Angle	0.74	0.03	-	-
Head-on	3.06	<.0001	1.71	<.0001
Safety Equipments	-2.59	<.0001	-2.00	0.01
Curved	-1.17	0.0003	-	-
Daytime	-[a]	-	-0.67	0.05
Variance				
Head-on	1.02	0.03	-	-
μ_1	1.93	<.0001	0.70	<.0001
μ_2	4.01	<.0001	2.82	<.0001
$L^*(\beta)$	-442		-205	
ρ^2	0.09		0.07	
No. of observations	372		370	

[a] A variable is excluded due to statistically insignificance of the variable at a 90% significance level.

Alternate HOL models were also developed to analyze effects of partner vehicle types on car, light truck and heavy truck drivers' injury severity separately as shown in

Table 5-5(a). The effects of explanatory variables on injury severity in these models are generally similar to the results of the previous two-vehicle crash models. The only difference is that these alternative models can capture the effect of partner vehicle types using dummy variables. The base case in each model is the collision between the same vehicle types (C-C, L-L or H-H). The result shows that driver's injury severity is higher when the partner vehicle is larger and heavier.

It should be noted that the goodness-of-fit is slightly better for these joint models (Table 5-5(a)) than the separate models (Tables 5-2~5-4) as indicated by higher values of log-likelihood ratio index (ρ^2). This is expected because of a larger sample size. However, these joint models can overlook the differences in effects of the same variable on driver's injury severity among different types of crashes, which have been identified from the comparison of the separate models.

Based on the results of HOL models, marginal effects of these dummy variables were also estimated as shown in Table 5-5 (b). The result shows that the collisions with smaller and lighter vehicles increase the probability of no injury but the collisions with higher and heavier vehicles increase the probabilities of fatal/major, minor and minimal injuries. As expected, the highest positive marginal effect on fatal/major injury was observed for C-H(C) followed by L-H(L) and C-L(C). This verifies that larger difference in size and weight between two vehicles involved in collisions causes more severe damages to a smaller and lighter vehicle.

Table 5-5. Comparison of Two-vehicle Crashes Among Car, Light Truck and Heavy Truck Drivers

(a) Parameters of HOL model

	Car drivers only		Light truck drivers only		Heavy truck drivers only	
Parameter	Estimate	Pr > t	Estimate	Pr > t	Estimate	Pr > t
Intercept	2.55	<.0001	1.54	<.0001	2.61	<.0001
Female	0.44	<.0001	0.37	<.0001	-	-
Young	-0.08	0.02	-	-	-	-
Daytime	-0.18	<.0001	-	-	-	-
Safety equipments	-1.94	<.0001	-1.47	<.0001	-1.78	<.0001
Abnormal condition	0.15	0.0001	-	-	-	-
Angle	0.43	<.0001	0.14	0.02	-	-
Head-on	1.61	<.0001	0.59	<.0001	1.24	<.0001
Sideswipe	0.20	<.0001	-	-	-	-
Asphalt over concrete	-0.18	<.0001	-	-	-	-
Median width (m)	-0.01	0.0002	-	-	-	-
Surface width (m)	-0.01	<.0001	-	-	-	-
Improper action	- [a]	-	-0.48	<.0001	-0.55	0.0002
Vehicle age	-	-	0.01	0.03	-	-
Weekend	-	-	-0.10	0.01	-	-
Undivided	-	-	0.10	0.01	-	-
Collision with light truck	**0.70**	<.0001		N/A	**-2.53**	<.0001
Collision with heavy truck	**1.28**	<.0001	**0.59**	<.0001		N/A
Collision with car		N/A[b]	**-0.19**	0.002	**-3.00**	<.0001
Variance						
Female	0.24	<.0001	-0.20	0.039	-	-
Head-on	1.00	<.0001	0.68	<.0001	-	-
Safety equipments	-	-	-1.37	0.001	-	-
Collision with light truck	-0.31	<.0001		N/A	-	-
Collision with heavy truck	-0.62	<.0001	-	-		N/A
Collision with car		N/A	-	-	-	-
μ_1	0.99	<.0001	0.48	<.0001	0.65	<.0001
μ_2	3.09	<.0001	1.69	<.0001	2.63	<.0001
$L^*(\beta)$	-11252		-4316		-1116	
ρ^2	0.11		0.09		0.13	
No. of observations	10412		4139		2297	

[a] A variable is excluded due to statistically insignificance of the variable at a 90% significance level.
[b] A dummy variable is excluded as collisions between same vehicle types are set to the base case.

(b) Marginal effects of partner vehicle types on driver's injury severity

	Car drivers (compared to C-C crashes)		Light truck drivers (compared to L-L crashes)		Heavy truck drivers (compared to H-H crashes)	
	Collision with		Collision with		Collision with	
Injury severity	Light truck	Heavy truck	Car	Heavy truck	Car	Light truck
No injury	-0.16	-0.30	0.08	-0.25	0.31	0.26
Minimal injury	0.01	0.02	-0.02	0.05	-0.10	-0.08
Minor injury	0.12	0.21	-0.05	0.16	-0.16	-0.13
Fatal	0.03	0.06	-0.01	0.04	-0.05	-0.04

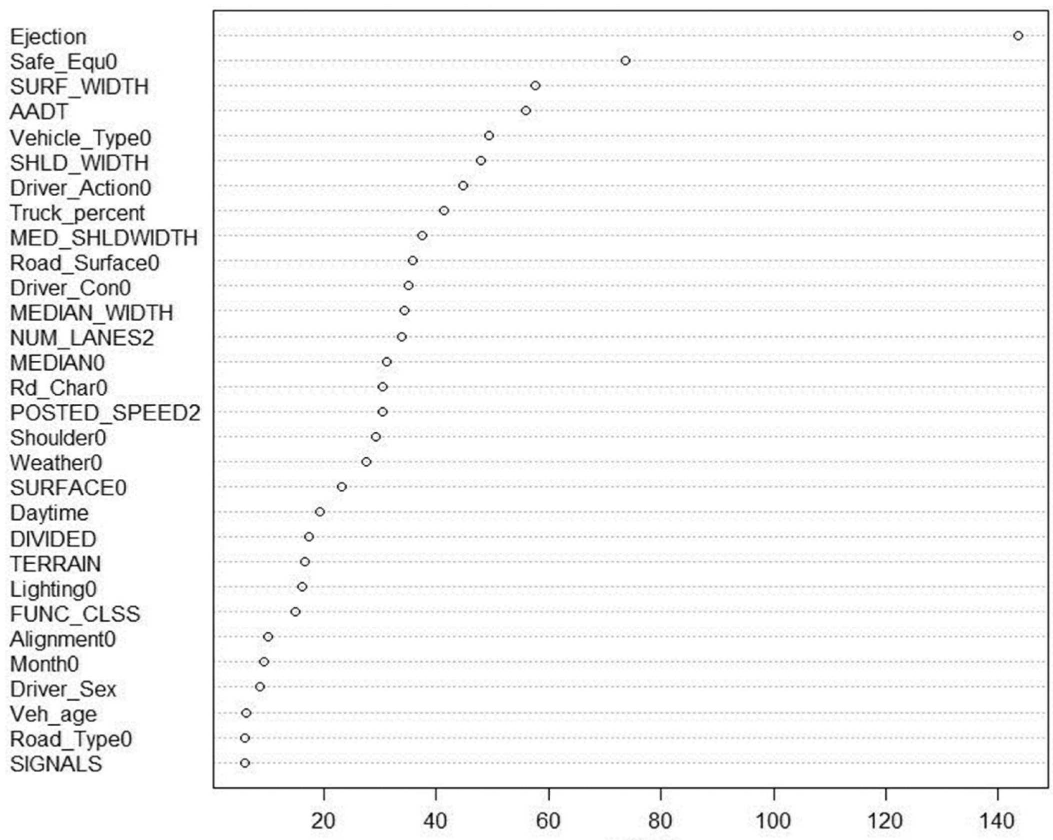
5.2. Random Forests

Ranking of importance of variables in prediction of driver's injury severity was determined using the random forests method for each crash type. The random forests method was applied with 500 trees and 2 randomly sampled candidate variables. Different numbers of trees were also considered but more than 500 trees significantly increased computation time with a minimal change in the results. Because the BRT model can only consider binary responses, the target variable (i.e. injury severity) was categorized into two levels - severe injury or non-severe injury – instead of four levels for consistency in classification of different models. The rankings of important variables for single-vehicle and car-car crashes are shown in Figure 5-1. The rankings for the other 8 crash types are shown in Appendix A.

In general, many variables show strong effects on injury severity. For example, ejection from vehicles, safety equipment, shoulder width, AADT and vehicle type were important for injury severity in single-vehicle crashes. In two-vehicle crashes, collision type (Impact), AADT, driver action, ejection and some road geometric variables had a significant influence on injury severity. For example, in car-car crashes, number of lanes, surface width and median width had strong effects on injury severity.

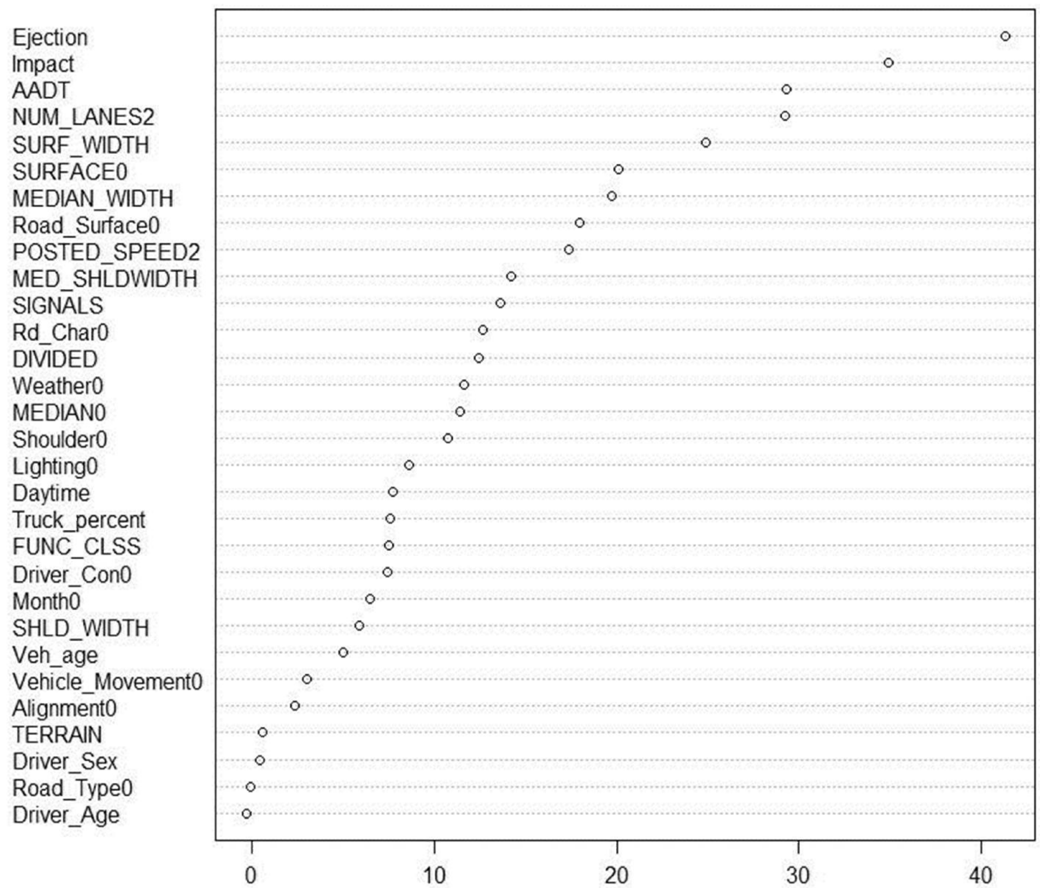
Other variables such as driver action and condition, driver's sex and truck percentage were also important factors affecting driver's injury severity in two-vehicle crashes. For C-H crashes, collision type had a higher importance ranking for car drivers than heavy truck drivers. Similar results were found for C-L crashes. This is because a smaller vehicle's driver is more likely to be impacted by collisions than a larger vehicle's driver in two-vehicle crashes.

It is worth to note that many road geometric and traffic variables – e.g. AADT, truck percentage, surface width, median width - were identified as important variables in the random forests method unlike HOL models. This is potentially because the random forest method which does not assume monotonic effects of explanatory variables on injury severity is more effective in reflecting their effects. These variables will be further investigated using the other non-parametric models – CART and BRT.



(a) Single-vehicle crashes

Figure 5-1. Rankings of Important Variables in Random Forests



(b) C-C crashes

Figure 5-1. Rankings of Important Variables in Random Forests (Continued)

However, although the random forests method can identify important variables, it is difficult to judge how injury severity will change as the value (or category) of the variable changes using the method unlike HOL models. Moreover, it is hard to capture non-linear effects of continuous variables on injury severity. These effects are likely to be more complex for continuous variables than binary and category variables.

5.3. Classification and Regression Trees

The CART was also applied to predict driver's injury severity for 10 crash types. The depth of tree was specified based on the sample size for crash type. Higher depth of tree was used for the crash type with a larger sample size to identify more split variables than the crash type with a smaller sample size.

Similar to the random forests method, injury severity was set as a binary variable - severe injury (= 1) or non-severe injury (= 0). Each node in the tree contains the numbers and percentages of the drivers who had severe injury and non-severe injury in both training data and validation data. From the trees, the factors contributing to driver's injury severity (i.e. split variables) were identified. Also, the effects of each split variable on severe injury were examined based on the percentages of severe injury in the training data set because the tree was developed using the training data set only.

5.3.1. Single-Vehicle Crash Model

The CART at the first level for single-vehicle crashes is shown in Figure 5-2. The full tree structure of the CART is shown in Figure B-1 in Appendix B. It was found that ejection from vehicles was the most important variable as it was the first split variable. The tree in Figure 5-2 shows that Node 1 (total drivers) was split into two nodes – Node 2 (ejected drivers) and Node 3 (non-ejected drivers). It was found that the proportion of severe injury within each node was higher for the ejected drivers (42.7%) than non-ejected drivers (6.2%). This indicates that drivers are more likely to be severely injured if they are ejected from vehicles.

The tree was split by ‘driver condition’ in the second level. If the driver was ejected, the proportion of severe injury was higher for abnormal driver condition (59.5%) than normal driver condition (29.1%). A similar trend was observed for non-ejected drivers.

In the third level of the tree, driver injury severity was higher for female drivers and non-use of safety equipment than male drivers and non-use of safety equipment, respectively. It was also found that the proportion of severe injury was higher for lower AADT (< 33,250) than higher AADT (≥ 33,250). The injury severity was higher for dry surface and daytime than the other surface conditions and nighttime, respectively.

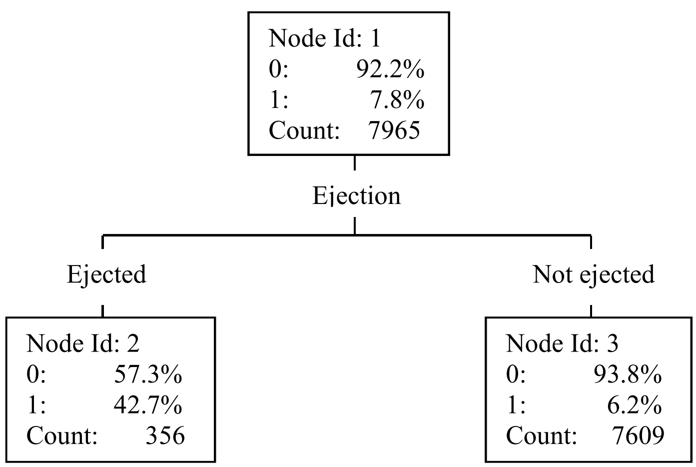


Figure 5-2. CART at the first level for Single-Vehicle Crashes

5.3.2. Two-Vehicle Crash Models

Since there are many cases of two-vehicle crashes, the results of the CART models for only C-C and C-L(L) crashes are discussed for demonstration purposes. The results of the

CART models for all types of two-vehicle crashes are shown in Figures B-2 to B-10 in Appendix B.

Car-Car Crash Model

Figure B-2 shows the CART for the car-car crashes. It was found that collision type was the first split variable. The proportion of fatal/major injury was higher for head-on collisions (29.2%) than the other types of collision (1.8%).

Driver condition and ejection from vehicles were the split variables in the second level. Similar to the single-vehicle crash model, abnormal driver conditions and ejection from vehicles lead to more severe injury.

For the drivers involved in head-on collisions, injury severity is associated with alignment, number of lanes and driver's age. Based on the proportion of fatal/major injury, injury severity was higher for curved roads, 4 or more lanes, and driver's age older than 64 years.

Among non head-on collision types, turning, sideswipe and angle collisions lead to higher injury severity than the other collision types (e.g. rear-end collisions). For these other collision types, nighttime, shoulder wider than or equal to 3.75 m and vehicles older than 18 years were associated with more severe driver injury severity. This indicates that drivers are more likely to make judgment errors on the roads with wider shoulder at nighttime. The result also shows that older vehicle model increases injury severity because newer vehicle models are usually equipped with better driver protection facilities that older vehicle models.

C-L(L) Crash Model

Figure B-7 shows the CART for the light truck drivers involved in car-light truck crashes. Similar to car-car crashes, head-on collisions, abnormal driver conditions, and ejection from vehicles increased injury severity. Flat terrain, female drivers, summer and arterials were also associated with fatal/major injury.

However, AADT had mixed effects on injury severity. In the third level, the proportion of fatal/major injury was higher for lower AADT (< 11,350) than higher AADT (≥ 11,350). But in the fourth level, the proportion of fatal/major injury was higher for higher AADT (≥ 24,850) than lower AADT (< 24,850). This indicates that the effect of AADT on light-truck driver's injury severity is nonlinear.

H-H Crash Model

Figure B-5 shows the CART for heavy truck-heavy truck crashes. It was found that truck percentage was the first split variable unlike the other types of two-vehicle crash. The percentage of severe injury was significantly higher for truck percentage greater than or equal to 32.95% (= 47.1%) than truck percentage less than 32.95% (= 0.4%). This indicates that higher truck percentage contributes more to heavy truck driver's severe injury. Similarly, shoulder width other than 3 and 3.5 m, clear weather and drivers younger than 42 increased the probability of severe injury.

5.3.3. Summary of Results in CART

The split variables in all CART model and their effects were summarized in Table 5-6. It was found that important factors associated with severe injury were different in different crash types. However, head-on collisions generally increased driver injury severity in two-vehicle crashes. The variable is also highly important since it was the first split variable for most crash types. Angle collisions also increased car driver's injury severity in C-H crashes. Similar to HOL models, ejection from vehicle, abnormal driving condition, female driver and vehicle age increased injury severity. Some environmental factors such as dark lighting condition and clear weather condition also increased injury severity. Moreover, some road geometric variables including asphalt pavement and number of lanes were associated with injury severity.

However, some continuous variables such as driver's age, shoulder width and median width had opposite effects among different crash types. These inconsistencies in the effects are because the CART splits the values of a continuous variable into two groups based on a single cut-off value and the model cannot clearly capture nonlinear effects of continuous variables.

Table 5-6. Important Split Variables and Their Effects on Severe Injury in CART for Single-vehicle and Two-vehicle Crashes

Variable	Single	C-C	C-H(C)	C-H(H)	H-H	C-L(C)	C-L(L)	L-L	L-H(L)	L-H(H)
Crash characteristics										
Head-on collision		+1	+1	+3		+1	+1	+1	+2	
Angle collision			+1							
February							+4			
Daytime	+4	+4								
Driver characteristics										
Ejected driver	+1	+2	+4				+2		+1	
Abnormal driver condition	+2	+2	+2				+2	+3		
Female driver	+3						+3			
Driver age*		+4			-3	+3				
Use safety equipment	-3									
Improper driver action			+2					+4		
Environmental characteristics										
Dry road surface	+4									
Dark condition			+							+1
Clear weather					+2			+2		
Vehicle characteristics										
Vehicle age*		+5				+3				
Geometric characteristics										
Curved road		+3						+4		
No. of lanes*		+3								
Shoulder width*		+5			-2	-5				
Flat terrain							+4			
Undivided road			+3							
1-m median shoulder width*						+4				
Asphalt pavement				-1			-4			
Median width*				+2						
Arterial										
Speed limit						-2				
Traffic characteristics										
AADT*	-3					-4	+/-3**			
Truck percentage*					+1	+2				

Note:
+: Positive effect, -: Negative effect
Number denotes the ranking of split or the level of importance (e.g. "1" denotes that the variable is the first split variable and it is the most important).

*Trend is unclear for continuous variables.
** Mixed effects.

5.4. Boosted Regression Trees

The BRT was also applied to predict driver's injury severity for 10 crash types using two-level injury severity. The BRT could not be developed for the C-H (H) case due to a low number of severely injured heavy truck drivers in car-heavy truck crashes. The marginal effects of the 12 most influential explanatory variables were estimated for each crash type with a learning rate 0.001 and tree complexity of five.

5.4.1. Single-Vehicle Crash Model

Figure 5-3 exemplifies the 12 most important variables for single-vehicle crashes. In general, the important variables were similar to those variables identified by the random forests method and the CART.

It was found that ejection from vehicles was the most important factor and it had a positive effect on severe injury for single-vehicle crashes. The other categorical factors including abnormal driver condition, improper driver action, dry road surface, male drivers and curved segments also had positive effects on severe injury. The figure also shows that the BRT can identify important nonlinear relationships between continuous variables and injury severity. For instance, although the marginal effect generally increased with driver's age, it abruptly increased for the drivers older than 75. This indicates that very old driver's risk of severe injury is significantly higher than younger driver's risk in single-vehicle crashes.

Similarly, marginal effects noticeably increased for truck percentage higher than 35% and vehicle age older than 25 years. Positive effect of older vehicles on higher injury severity was also reported in Kim et al. (2012). These nonlinear effects would have not been captured in parametric models which conventionally define these variables as a continuous linear predictor. Thus, the BRT is more advantageous in identifying nonlinear effects of variables than parametric models.

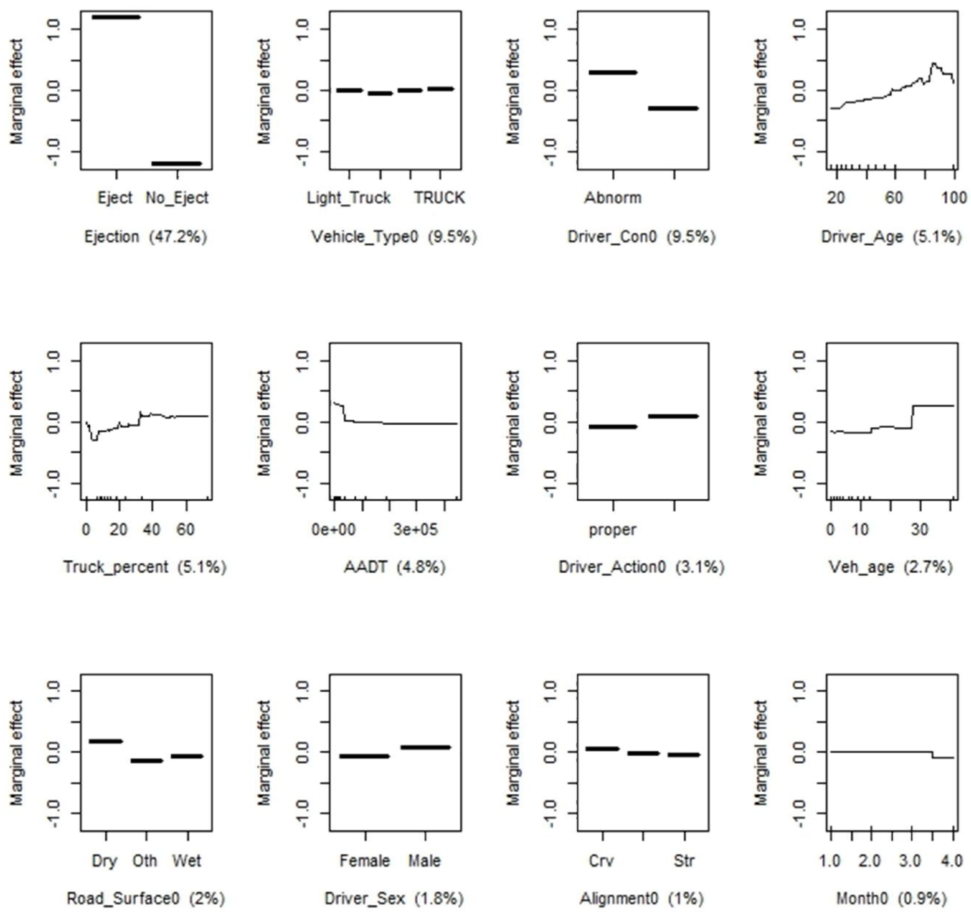


Figure 5-3. Marginal Effects of the 12 Most Important Variables for Single-Vehicle Crashes in BRT

5.4.2. Two-Vehicle Crash Models

Although there are 9 cases of two-vehicle crashes, the results of the BRT models for only L-H crashes are discussed for demonstration purposes. The results of the BRT models for the other crash types are shown in Appendix C.

Figures 5-4 and 5-5 show the marginal effects for the L-H models. For example, in L-H(H) model, injury severity increases sharply when AADT is higher than 300,000. However, in L-H(L) model, the effect of AADT decreases in the beginning and follows with a steady line. This indicates that heavy truck drivers may be severe injured when crash occurs on a high traffic volume (AADT) road. Moreover, the effects of impact are different. Relative effect of head-on collisions compared to angle collisions was higher for heavy truck driver's injury than light truck driver's injury. This implies that heavy truck drivers are relatively safer than light truck drivers in angle collisions. This is probably because the impact of angle collisions on a larger vehicle's driver is lower than the impact on a smaller vehicle's driver.

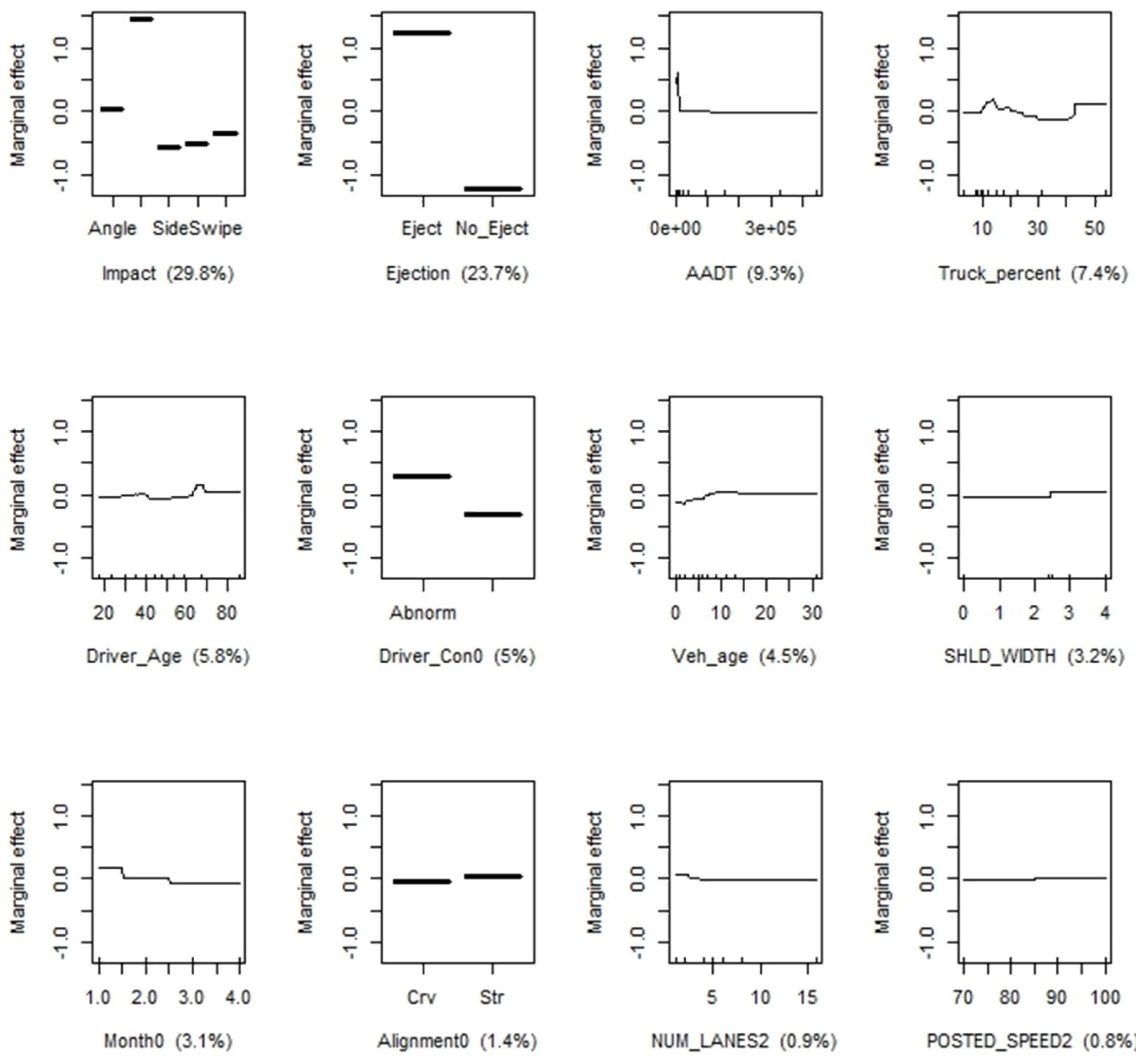


Figure 5-4. Marginal Effects of the 12 Most Important Variables for L-H(L) Crashes in BRT

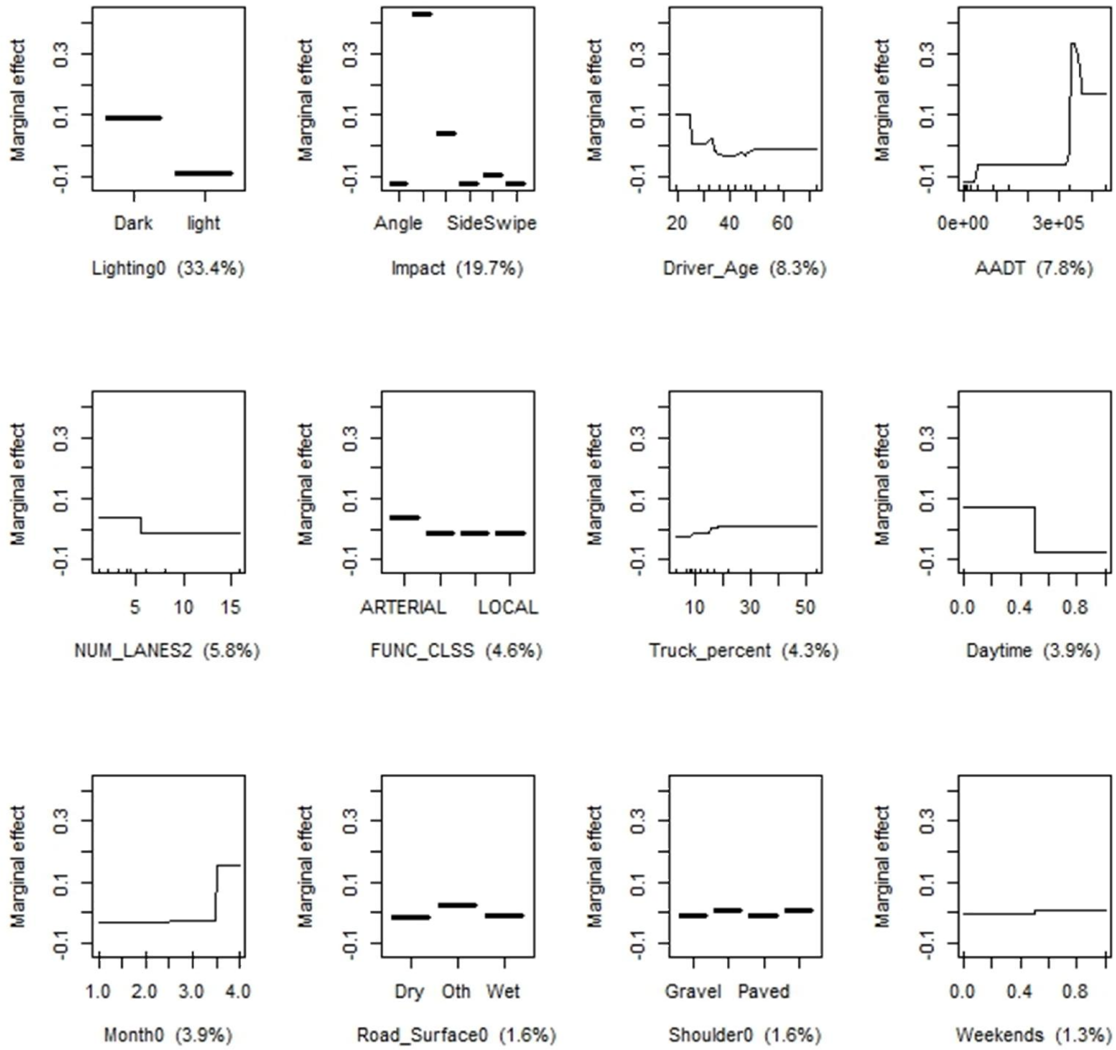


Figure 5-5. Marginal Effects of the 12 Most Important Variables for L-H(H) Crashes in BRT

5.4.3. Summary of Results in BRT

As mentioned in previous chapter, non-linear effects are hard to be explored by the CART and HOL models. The effects of some variables may have nonlinear effects on driver injury severity. This is why the effects of variables were not consistent in different studies. For example, some studies found that younger drivers are more likely to be severe injured (Harb et al., 2008; Weiss et al., 2014) but the other studies claimed that older drivers are more likely to be severely injured (Zhang et al., 2000; Kim et al., 2012).

Table 5-7 summarizes the 12 most important variables for 9 crash types. It is worth to note that ejection, AADT, truck percentage, driver's age and vehicle age (except L-H(H) crashes) were important in almost all crash types. Collision type was also commonly important for two-vehicle crashes. Based on the plots of marginal effects, the effects of these six variables on severe injury were discussed as follows.

Table 5-7. Important Variables for Single-vehicle and Two-vehicle Crashes in BRT

Single-vehicle	C-C	C-H(C)	H-H	C-L(C)
Ejection	**Collision type**	**Collision type**	**Truck percentage**	**Collision type**
Vehicle type	**Ejection**	**Ejection**	Shoulder width	**Truck percentage**
Driver condition	**Vehicle age**	Driver action	Driver condition	**AADT**
Driver age	**AADT**	**AADT**	**Vehicle age**	**Vehicle age**
Truck percentage	**Driver age**	**Driver age**	Median width	**Driver age**
AADT	**Truck percentage**	**Truck percentage**	**Driver age**	**Ejection**
Driver action	Road surface	Driver condition	Surface condition	Shoulder width
Vehicle age	Posted speed limit	Lighting	Month	Month
Road surface	Surface width	Surface width	**AADT**	Posted speed limit
Driver sex	Road type	Month	Pavement material	Road type
Alignment	Month	**Vehicle age**	Surface width	Surface width
Month	Shoulder type	Shoulder width	Time of day	Surface condition

C-L(L)	L-L	L-H(L)	L-H(H)
Collision type	**Collision type**	**Collision type**	Lighting
AADT	**AADT**	**Ejection**	**Collision type**
Ejection	**Truck percentage**	**AADT**	**Driver age**
Driver age	**Driver age**	**Truck percentage**	**AADT**
Vehicle age	Road surface	**Driver age**	Number of lanes
Truck percentage	**Ejection**	Driver condition	Road type
Surface condition	Shoulder width	**Vehicle age**	**Truck percentage**
Surface width	**Vehicle age**	Shoulder width	Time of day
Driver condition	Weather	Month	Month
Month	Shoulder type	Alignment	Surface condition
Shoulder type	Surface width	Number of lanes	Shoulder type
Median width	Alignment	Posted speed limits	Day of week

The effects of binary or categorical variables on severe injury were consistent in all crash types. As expected, it was found that severe injury is most likely to occur when drivers are ejected from vehicles and they are involved in head-on collisions (in case of two-vehicle crashes). These results are similar to the HOL and CART models. It was also found that the marginal effect of angle crashes was relatively lower for heavy truck drivers in L-H crashes compared to the other crash types. This indicates that heavy truck drivers are less likely to be severely injured if they are involved in angle collisions with smaller vehicles.

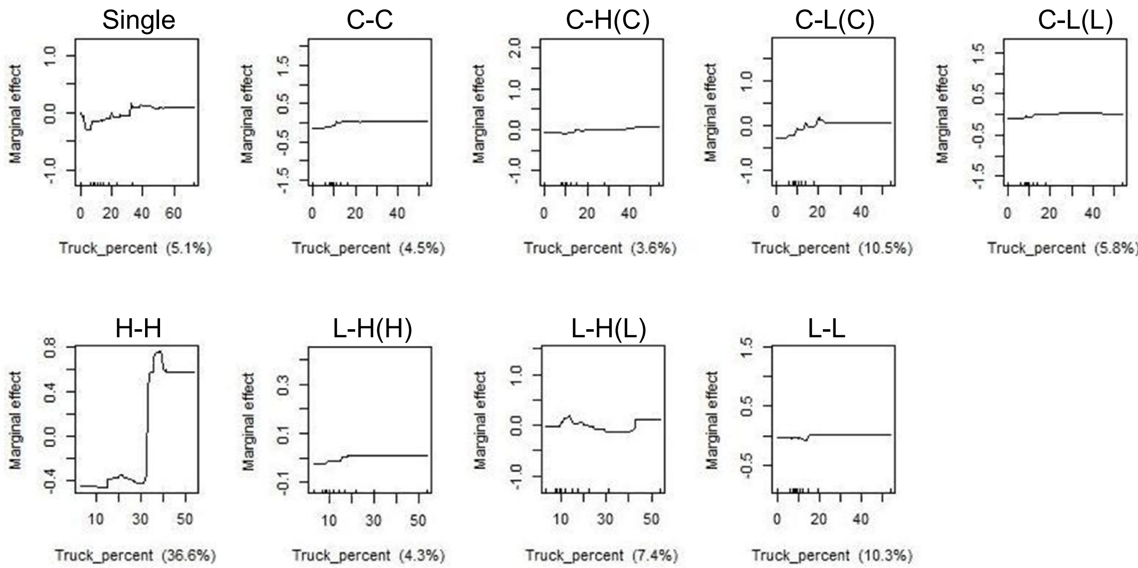
However, the effects of continuous variables on severe injury were not consistent among different crash types. For instance, the marginal effect of truck percentage on severe injury was different for H-H crashes compared to the other crash types as shown in Figure 5-6(a) - the effect sharply increased as truck percentage exceeded 30%. This indicates that higher truck percentage is more likely to increase the chance of heavy truck driver's severe injury in H-H crashes.

It was also found that the marginal effect of AADT generally decreases as AADT increases similar to Duncan et al. (1998) except heavy truck drivers involved in L-H crashes and H-H crashes (Figure 5-6(b)). For this crash type, injury severity increased with AADT unlike the other crash types where injury severity generally decreased with AADT. A similar but weaker trend was also observed for H-H crashes. Higher truck percentage and higher AADT reflect more frequent interactions among vehicles and more complex driving environments – e.g. higher speed variation, more frequent lane changes, etc. It appears that heavy truck drivers are more likely to make judgment errors and they are more severely injured in such traffic condition.

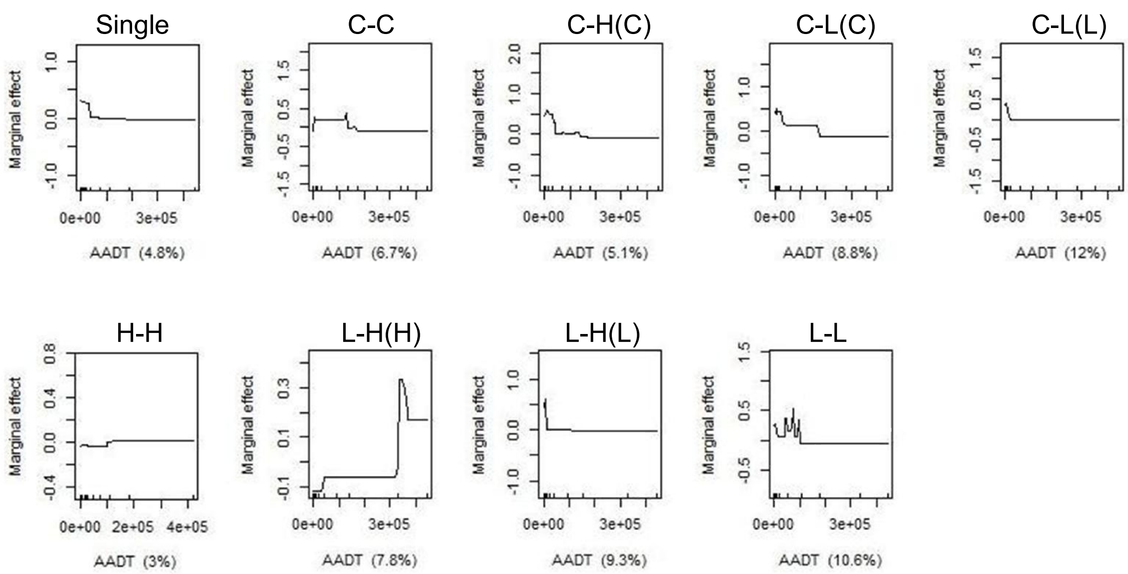
In case of driver's age, the pattern of change in the marginal effect was significantly different for heavy truck drivers in L-H crashes as shown in Figure 5-6(c). For this crash type, injury severity decreased as driver's age increased. A similar but weaker trend was also observed for H-H crashes. This indicates that younger heavy truck drivers are more likely to be severely injured than older heavy truck drivers. This is potentially because younger heavy truck drivers are less experienced than older drivers.

However, marginal effects of vehicle age were almost similar for all crash types as shown in Figure 5-6(d). In general, driver's severe injury is more likely to occur for older vehicles. This is mainly because older vehicle models have relatively fewer safety features than new vehicle models.

The results show that capturing these non-linear effects is the biggest advantage of the BRT model. However, the BRT model cannot quantify these non-linear effects. Thus, it is recommended that nonlinear effects of continuous variables are identified using the BRT models and then the variables are categorized to reflect the nonlinear effects. Then these categorical variables are included in the HOL model to estimate their quantitative effects.

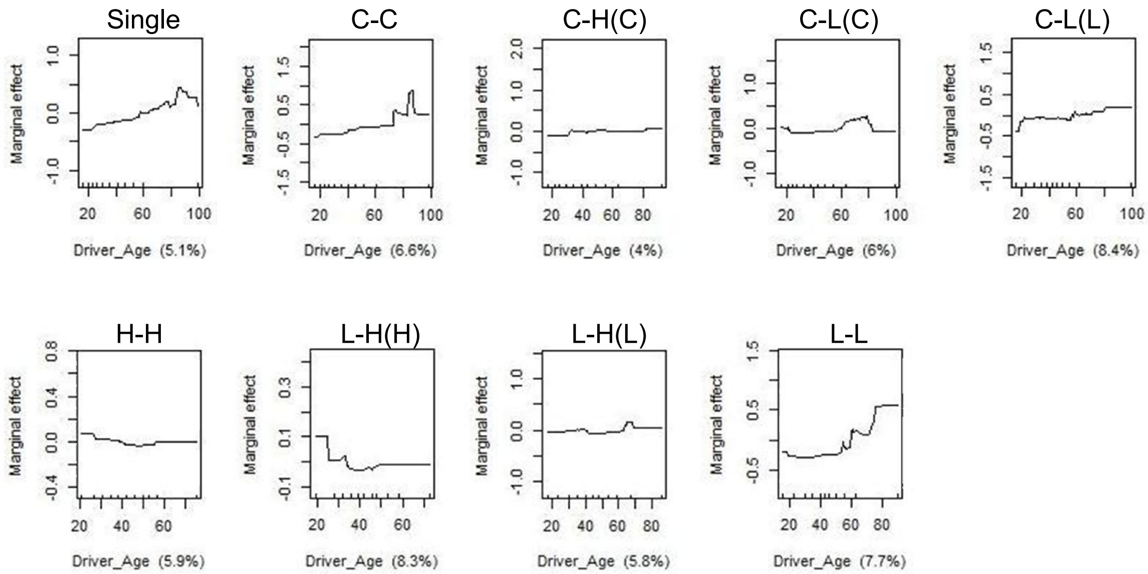


(a) truck percentage

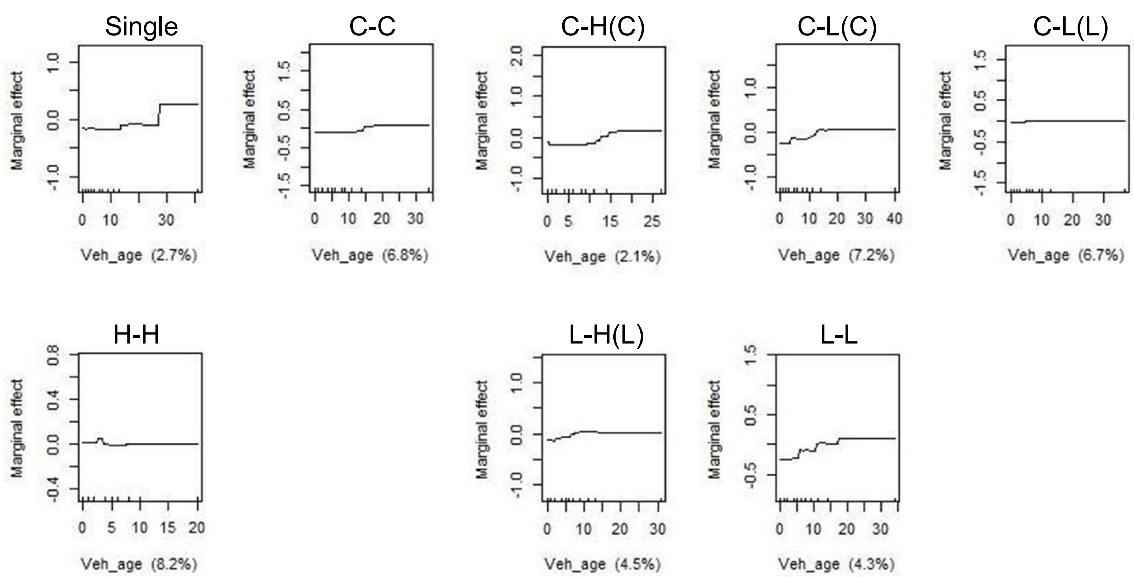


(b) AADT

Figure 5-6. Comparison of Marginal Effects of Continuous Variables for Different Crash Types in BRT.



(c) Driver's age



(d) Vehicle age

Figure 5-6. Comparison of Marginal Effects of Continuous Variables for Different Crash Types in BRT. (Continued)

5.5. Evaluation of Model Performance

In this section, the performance of the HOL, CART and BRT models was evaluated based on their classification accuracy. The HOL models were re-developed using two levels of injury severity (instead of four levels) for 10 crash types to be consistent with the CART and BRT models. The results of the HOL models with 2-level injury severity are presented and discussed in Appendix D.

The prediction of the HOL and BRT models is the probability that a driver's injury level is severe rather than the category of injury severity. Thus, the predicted category of injury severity was determined based the probability and a selected cut-off value (default = 0.5 for a binary response variable). In case of a default cut-off value, if the probability is greater than 0.5, the driver's injury severity is predicted as severe and vice versa.

However, due to very low proportion of severe injury compared to non-severe injury, most predicted injury severity is likely to be non-severe injury if a default cut-off value is used. Thus, to identify more severe injury correctly, a cut-off value should be decreased. But this will also increase the number of incorrectly classified severe injury. Thus, the cut-off value should be determined based on the following four cases:

True positive: the driver's injury severity is severe and the prediction is severe.

False positive: the driver's injury severity is not severe but the prediction is severe.

True negative: the driver's injury severity is not severe and the prediction is not severe.

False negative: the driver's injury severity is severe but the prediction is not severe.

Using the above cases, the sensitivity and the specificity are defined as the capability of the model to correctly identify the driver's severe injury and non-severe injury, respectively, as follows (Lalkhen and McCluskey, 2008).

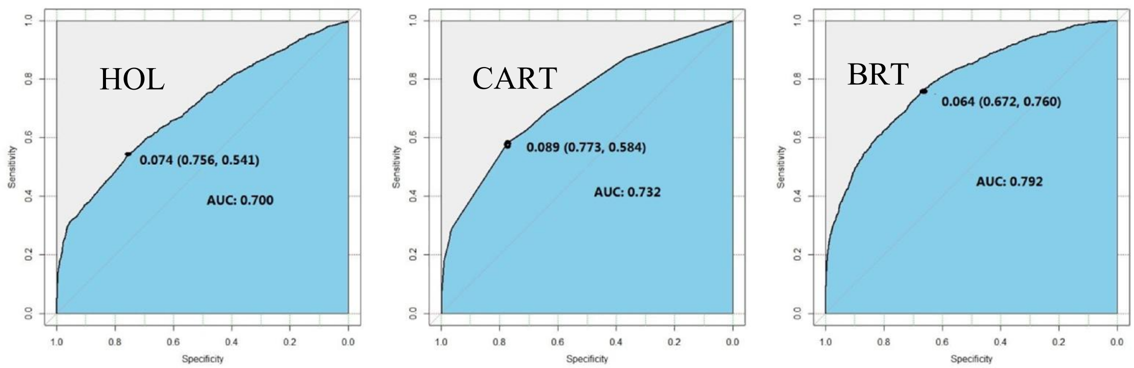
$$\text{Sensitivity} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$$

$$\text{Specificity} = \frac{\text{True negatives}}{\text{True negatives} + \text{False positives}}$$

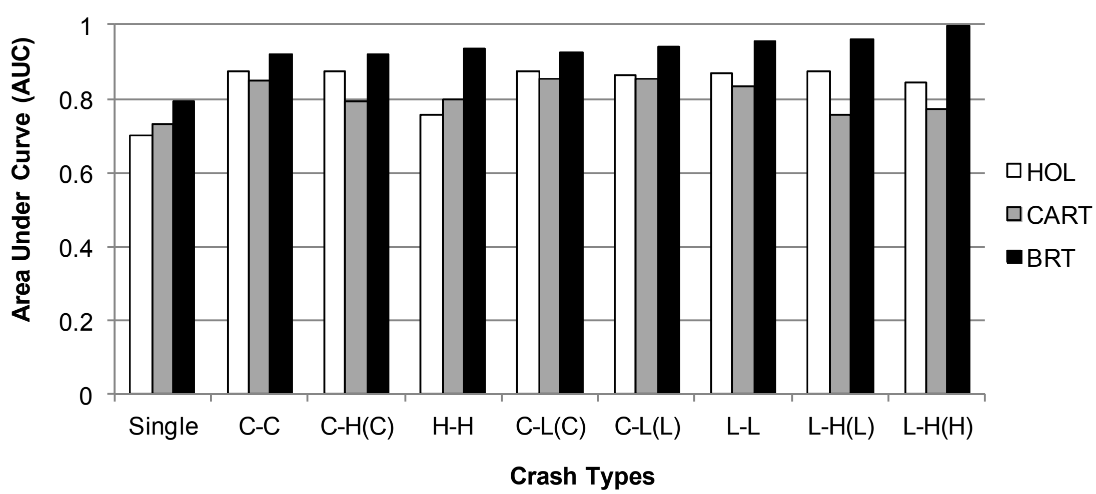
The model with higher classification accuracy will show higher sensitivity and specificity. However, there is a trade-off between the sensitivity and specificity. If the cut-off value is decreased to increase the sensitivity (i.e. correctly identify severe injury), the specificity (i.e. correctly identify non-severe injury) will decrease, and vice versa (Lalkhen and McCluskey, 2008). Thus, the optimal cut-off value should be determined such that the sensitivity and the specificity are balanced.

The relationship between the sensitivity and the specificity is graphically described in the Receiver Operator Characteristic (ROC) curve. The curve is drawn using different values of the sensitivity and the specificity for different cut-off values. A larger area under the curve (AUC) represents higher classification accuracy of the model. The AUC varies between 0 and 1. For demonstration purposes, Figure 5-7(a) shows the ROC curve and the AUC for single-vehicle crashes. The ROC curves for two-vehicle crashes are shown in Appendix E. It was found that AUC's were consistently larger for the BRT model than the HOL and CART models for all 9 crash types as shown in Figure 5-7(b).

This indicates that the BRT model can better predict driver's severe injury than the HOL and CART models.



(a) ROC curves for single-vehicle crashes



(b) Area under ROC curves for each crash type

Figure 5-7. Comparison of Goodness-of-fit among HOL, CART and BRT using ROC Curves

Note: The numbers on the ROC curve denote the optimal cut-off value with corresponding specificity and sensitivity in parenthesis.

6. Conclusions and Recommendations

This study applied both parametric and non-parametric models to identify the factors affecting injury severity of drivers involved in crashes and analyze the effects of the factors on injury severity. Many factors such as crash, driver, vehicle, traffic, environmental, and road geometric characteristics were examined. To consider the difference in weights of vehicles and impact of collisions on vehicle bodies, vehicles were classified into passenger car, light truck and heavy truck. Separate models were developed for single-vehicle crashes and two-vehicle crashes classified by different combinations of vehicle types.

Among many parametric models, the heteroscedastic ordered logit (HOL) model was used because it can account for variation in the unobserved effects of variables among observations unlike conventional ordered logit model. For non-parametric models, the boosted regression trees (BRT) model was used because it fits multiple trees and it can more accurately classify the cases which are more difficult to be classified unlike conventional classification and regression tree (CART) model. Next, the Receiver Operator Characteristic (ROC) curves method was used to evaluate the prediction accuracy of each model. The findings in this study are summarized as follows:

1. In all models, it was commonly found that some factors influence driver's injury severity for both single and two-vehicle crashes. As expected, driver's ejection from vehicles and driver's age increased injury severity. Collision type was the most significant variable in two-vehicle crashes. Head-on and angle collisions were the most dangerous crashes for passenger car and light truck drivers.

2. Some variables had varying effects in different types of crashes in the HOL and BRT models. For example, young driver's (≤ 30) injury severity increased in single-vehicle crashes but it decreased in car-car crashes in the HOL models. On the other hand, marginal effect of driver's age on severe injury was opposite between heavy truck drivers and car/light truck drivers – injury severity increased as heavy truck driver's age decreased.

3. The HOL model can capture the variation in the effects of some variables among different drivers and crashes. In particular, the variations in driver's injury severity were significant in head-on collisions between two vehicles. This indicates that injury severity of drivers involved in head-on collisions highly depends on the other factors such as point of impact and collision force.

4. A smaller and lighter vehicle's drivers are more likely to be severely injured when they are involved in a collision with a larger and heavier vehicle. In particular, the probability of fatal and major injury is higher when the difference in vehicle size and weight between two vehicles is greater.

5. The BRT model can capture nonlinear effects of variables without pre-specified relationship between variables and injury severity. The plots of marginal effects showed that some continuous variables including road geometric and traffic factors had nonlinear effects on severe injury.

6. Traffic factors were significant in only non-parametric models, but not in the case of parametric models. More specifically, AADT and truck percentage had strong effects in the BRT and CART models, but they were not significant in the HOL models. This is potentially because the effects of these continuous variables are

more complex than the effects of categorical variables and there exist interactions or correlations among these variables. This indicates that these effects can be better captured by non-parametric models.

7. The BRT models showed better performance than the HOL and CART models for all crash types based on the comparison of area under the ROC curve. However, the HOL model showed better performance than the CART for most crash types.

The study demonstrates that separate models for single-vehicle and different types of two-vehicle crashes can identify differential effects of factors on driver's injury severity. Both parametric and non-parametric models generally identified similar factors affecting injury severity but they have advantages and disadvantages. Parametric models can estimate quantitative effects of variables based on coefficients for each parameter. However, since they assume pre-specified monotone relationships between injury severity and independent variables, it is difficult to capture nonlinear effects of certain variables. On the other hand, non-parametric models do not require pre-defined relationships and capture complex relationships better than parametric models. They can also avoid the problems of multi-collinearity among variables and outliers. However, they cannot estimate the quantitative effects of variables unlike nonparametric models. Thus, both parametric and nonparametric models are recommended for prediction of injury severity. For instance, important variables are identified using the BRT model and these variables are included in the HOL model to investigate their quantitative effects.

Based on the results in this study, some remedial treatments are suggested to reduce driver's injury severity. First, increasing median width and surface width and curvature could reduce driver's injury severity associated with car-car crashes and single-vehicle crashes, respectively. If a sufficient space for increasing road width is not available, some technological improvements can be performed – e.g. installation of crash cushions to reduce damage to vehicles from collisions. Second, educating and training heavy truck drivers and young drivers to take more caution in low traffic volume conditions where they are more likely to drive fast. Also, young heavy truck drivers are recommended to drive more cautiously on the roadways with high truck percentage and traffic volume. Third, a special design consideration is needed for undivided roadways with high truck volume to prevent head-on collisions between passenger cars and heavy trucks.

However, there are some limitations in this study. First, some important variables were missing in the data sets. For instance, the actual speed of vehicle prior to the crash was unknown. This variable is critically important since higher speed at the time of crash increases the impact of collision on drivers and lead to more severe injury. Although posted speed limits can reflect driver's average speed, they may not be the same as actual speed. Also, since the exact point of impact is unknown, it is still unclear why variance in injury severity is higher for certain collision types, particularly head-on collisions. Moreover, drivers' physical condition and driving experience were not available in the data although these factors are strongly associated with driver's injury severity and their driving habits, respectively. Second, due to a small sample size of heavy truck drivers, relatively less number of significant factors was identified for heavy-truck-involved crashes and heavy truck driver's injury in car-heavy truck crashes could not be analyzed using the BRT model. Third, interaction effects of multiple variables were not considered in the models using interaction terms.

In future study, it is recommended that more data should be collected to validate the results of the HOL and BRT models. For instance, two-vehicle crash models can be validated using the records of drivers' injury severity not selected in the model development. It is also recommended that the models are applied more extensively to predict injury severity in multi-vehicle crashes involving more than two vehicles. Lastly, various traffic control strategies need to be developed to separate or harmonize car and truck movements to minimize their conflicts and reduce risk of severe injury caused by collisions.

References

Abay, K. A., Paleti, R., Bhat, C. R., 2013. The joint analysis of injury severity of drivers in two-vehicle crashes accommodating seat belt use endogeneity. Presented at 92nd Transportation Research Board Annual Meeting.

Ahmed, M., Abdel-Aty, M., 2013. A data fusion framework for real-time risk assessment on freeways. Transportation Research Part C 26, pp. 203-213.

Aitchison, J., Silvey, S., 1957. The generalization of probit analysis to the case of multiple responses. Biometrika 44, pp. 131-140.

Anastasopoulos P.Ch, Mannering F.L, 2011. An empirical assessment of fixed and random parameter logit models using crahs-and non-crash-specific injury data. Accident Analysis and Prevention 43, pp. 1140-1147.

Anowar, S., Yasmin, S., Tay, R., 2012. Severity of single vehicle crashes during holidays. Presented at 91st Transportation Research Board Annual Meeting.

Bedard, M., Guyatt, G., Stones, M., Hirdes, J., 2002. The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. Accident Analysis and Prevention 34, pp. 717-727.

Carslaw, D. C., Taylor, P. J., 2009. Analysis of air pollution data at a mixeds Location using Boosted Regression Trees. Atmospheric Environment 43, pp. 3563–3570.

Chang, H.-L., Yeh, T.-H., 2007. Motorcyclist accident involvement by age, gender, and risky behaviors in Taipei, Taiwan. Transportation Research Part F, Vol. 10, Iss. 2, pp. 109-122.

Chang, H.-L., Chien J-T., 2013. Analysis of driver injury severity in truck-involved accidents using a non-parametric classification tree model. Accident Analysis and Prevention 51, pp. 17-22.

Chang, L.-Y., Wang, H.-W., 2006. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. Accident Analysis and Prevention 38, pp. 1019-1027.

Chen, F., Chen, S., 2011. Injury severities of truck drivers in single- and multi-vehicle accidents on rural highways. Accident Analysis and Prevention 43, pp. 1677-1688.

Cheong Y. L., Leitão, P. J., Lakes, T. Assessment of land use factors associated with dengue cases in Malaysia using Boosted Regression Trees. Spatial and Spatio-temporal Epidemiology, In press.

Chung, Y.-S., 2013. Factor complexity of crash occurrence: An empirical demonstration using boosted regression trees. Accident Analysis and Prevention 61, pp. 107-118.

Duncan, C. S., Khattak, A. J., Council, F. M., 1998. Applying the ordered probit model to injury severity in truck-passenger car rear-end collisions. Transportation Research Record, No 1653, pp. 63-71.

Elith, J., Leathwick, J., 2014. Boosted Regression Trees for Ecological Modeling. http://cran.r-project.org/web/packages/dismo/vignettes/brt.pdf.

Elith, J., Leathwick, J. R., Hastie, T., 2008. A working guide to boosted regression trees. Journal of Animal Ecology 77, pp. 802-813.

Friedman, J. H., 2001. Greedy function approximation: A gradient boosting machine. Annals of Statistics 29 (5), pp. 1189–1232.

Geedipally, S. R., 2014. A injury severity analysis of crashes on ramps and at crossroad ramp terminals. Presented at 93rd Annual Meeting of Transportation Research Board, Washington D.C.

Habib, M. A., Forbes, J. J., 2014. Modeling bicyclists' injurys levels in the Province of Nova Scotia, Canada using a Generalized Ordered Probit structure. Presented at 93rd Annual Meeting of Transportation Research Board, Washington D.C.

Harb, R., Yan, X., Radwan, E., Su, X., 2009. Exploring precrash maneuvers using classification trees and random forests. Accident Analysis and Prevention 41, pp. 98-107.

Hastie, T., Tibshirani, R., Friedman, J., 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition. Springer.

Ho, T., 1995. Random decision forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, Canada.

Huang, H., Chor, C., Haque, Md., 2008. Severity of driver injury and vehicle damage in traffic crashes intersections. Accident Analysis and Prevention 40, pp. 45-54.

Islam, M., Hernandez, S., 2014. Modeling injury outcomes of crashes involving heavy vehicles in rural and urban settings in Texas. Presented at 93rd Annual Meeting of Transportation Research Board, Washington D.C.

Jiang, X., Huang, B., Zaretzkib, R. L., Richards, S., Yan, X., Zhang, H., 2013. Investigating the influence of curbs on single-vehicle crash injury severity utilizing zero-inflated ordered probit models. Accident Analysis and Prevention 57, pp. 55-66.

Khattak, A. J., Schneider, R. J., Targa, F., 2003. Risk factors in large truck rollovers and injury severity: analysis of single-vehicle collisions. Presented at 82nd Transportation Research Board Annual Meeting.

Khorashadi, A., Niemeier, D., Shanker, V., Mannering F., 2005. Differences in rural and urban driver-injury severities in accidents involving large-trucks: An exploratory analysis. Accident Analysis and Prevention 37, pp. 910–921.

Kim, J.-K., Ulfarsson G. F., Shankar,V. N, Mannering, F.L., 2010. A note on modeling pedestrian-injury severity in motor-vehicle crashes with the mixed logit model. Accident Analysis and Prevention 42, pp. 1751-1758.

Kim, J.-K., Ulfarsson, G. F., Kim, S., Shankar, V. N., 2012. Driver-injury severity in single-vehicle crash in California: A mixed logit analysis of heterogeneity due to age and gender. Accident Analysis and Prevention 50, pp. 1073-1081.

Kockelman, K.M., Kweon, Y.-J., 2002. Driver injury severity: An application of ordered probit models. Accident Analysis and Prevention 34, pp. 313-321.

Lalkhen, A. G., McCluskey, A., 2008. Clinical tests: Sensitivity and specificity. Continuing Education in Anaesthesia, Critical Care & Pain, Vol. 8, No. 6, pp. 221-223.

Lee, C., Abdel-Aty, M., 2008. Presence of passengers: Does it increase or reduce driver's crash potential? Accident Analysis and Prevention 40, pp 1703-1712.

Lemp, J., Kockelman, K., Unnikrishnan, A., 2011. Analysis of large truck crash severity using heteroskedastic ordered probit models. Accident Analysis and Prevention 43, 370-380.

Ministry of Transportation Ontario, 2007. Analysis and Estimation of the Social Cost of Motor Vehicle Collisions in Ontario. http://www.tc.gc.ca/media/documents/ roadsafety/ TP14800E.pdf. Accessed Dec. 2012.

Ministry of Transportation Ontario, 2012. Ontario Road Safety Annual Report (ORSAR) 2010. http://www.mto.gov.on.ca/english/safety/orsar/orsar10/people.shtml. Accessed Dec. 2012.

Nassiri, H., Edrissi, A., 2006. Modeling truck accident severity on two-lane rural highways. Scientia Iranica, Vol.13, No.2, pp. 193-200.

Neyens, D. M., Boyle, L. N., 2008. The influence of driver distraction on the severity of injuries sustained by teenage drivers and their passengers. Accident Analysis and Prevention 40, pp. 254–259.

Pande, A., Abdel-Aty, M., Das, A., 2010, A classification tree based modeling approach for segment related crashes on multilane highways. Journal of Safety Research 41, pp. 391–397.

Peng Y., Boyle, L. N., 2012. Commercial driver factors in run-off-road crashes. Transportation Research Record: Journal of the Transportation Research Board, No. 2281, Transportation Research Board of the National Academies, Washington, D.C., pp. 128-132.

Qin, X., Wang, K., Cutler, C. E., 2013. Modeling large truck safety using logistic regression models. Presented at 92nd Transportation Research Board Annual Meeting.

R Development Core Team, 2006. The R Project for Statistical Computing.

Rana, T. A., Sikder, S. S., Pinjari, R., 2010. Copula-based method for addressing endogeneity in models of severity traffic crash injuries. Transportation Research Record: Journal of the Transportation Research Board, No. 2147, Transportation Research Board of the National Academies, Washington, D.C., pp. 75-87.

SAS Institute Inc., 2012. SAS 9.2. Cary, NC, USA.

SAS Institute Inc., 2009. SAS Enterprise Miner 6.2. Cary, NC, USA.

Savolainen, P. T., Mannering, F. L, 2007. Probabilistic models of motorcyclists' injury severities in single- and multi-vehicle crashes. Accident Analysis and Prevention 39, pp. 955-963.

Schneider IV, W. H., Savolainen, P. T., Zimmerman, K., 2009. Driver injury severity resulting from single-vehicle crashes along horizontal curves on rural two-lane highways. Transportation Research Record: Journal of the Transportation Research Board, No. 2102, Transportation Research Board of the National Academies, Washington, D.C., pp. 85-92.

Sobhani, A., Young, W., Logan, D., Bahrololoom, S., 2011. A kinetic energy model of two-vehicle crash injury severity. Accident Analysis and Prevention 43, pp. 741-754.

Srinivasan, K. K., 2002. Injury severity analysis with variable and correlated thresholds. Transportation Research Record 1784, pp. 132-142.

Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T., 2007. Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. BMC Bioinformatics 8(25).

Torrão, G. A., Coelho, M. C., Rouphail, N. M., 2014. Modeling the impact of subject and opponent vehicle on crash severity in two-vehicle collisions. Presented at 93rd Transportation Research Board Annual Meeting, Washington, D.C.

Transport Canada, 2010. Heavy Truck Casualty Collisions 2001 – 2005. TP 2436 E, Fact Sheet RS 2010-03. http://www.tc.gc.ca/eng/roadsafety/tp-tp2436-rs201003-1042.htm. Accessed Dec. 2012.

Transport Canada, 2012. Canadian Motor Vehicle Traffic Collision Statistics: 2010. http://www.tc.gc.ca/eng/roadsafety/tp-1317.htm. Accessed Dec. 2012.

Vodden, K., Smith, D., Eaton, F., Mayhew, D., 2007. Analysis and Estimation of the Social Cost of Motor Vehicle Collisions in Ontario, Final Report. http://www.tc.gc.ca/media/documents/roadsafety/tp14800e.pdf.

Wang X., Kockelman, K., 2005. Use of heteroscedastic ordered logit model to study severity of occupant injury: Distinguishing effects of vehicle weight and type. Transportation Research Record: Journal of the Transportation Research Board, No.

1908, Transportation Research of the National Academies, Washington, D.C., pp.195-204.

Wang, K., Qin, X., 2014. Using structural equation modeling to measure single-vehicle crash severity. Presented at 93rd Transportation Research Board Annual Meeting, Washington, D.C.

Wang, X, Abdel-Aty, M., 2008. Analysis of left-turn crash injry severity by conflicting pattern using partial proportional odds models. Accident Analysis and Prevention 40, pp. 1674-1682.

Weiss, H. B., Kaplan, S., Prato, C. G., 2014. Analysis of factors associated with injury severity in crashes involving young New Zealand drivers. Accident Analysis and Prevention 65, pp. 142–155.

Williamson, A., Lombardi, D. A., Folkard, S., Stutts, J., Courtney, T. K, Connor, J. L., 2011. The link between fatigue and safety. Accident Analysis and Prevention 43, pp. 498-515.

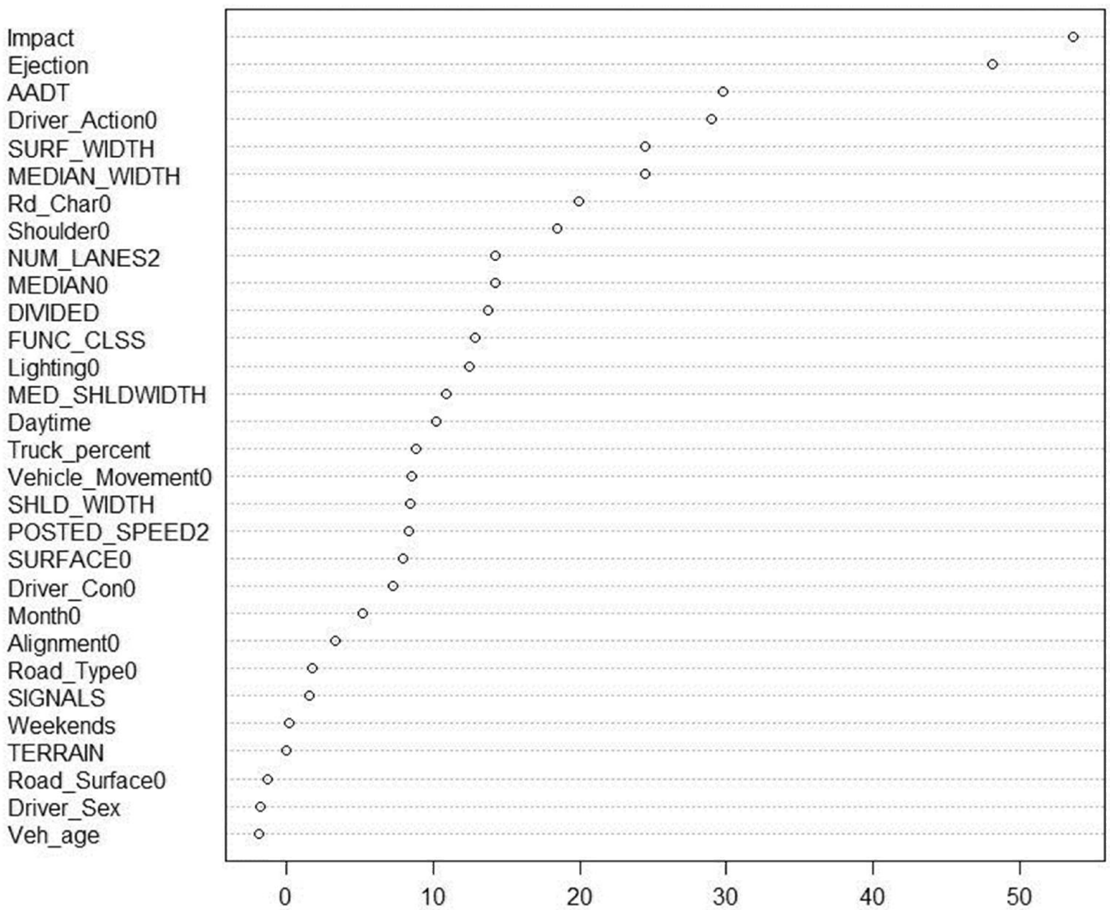
Xie, Y., Zhao K., Huynh N., 2012. Analysis of driver injury severity in rural single-vehicle crashes. Accident Analysis and Prevention 47, pp. 36-44.

Zajac, S., Ivan, J., 2003. Factors influencing injury severity of motor vehicle crossing pedestrian crashes in rural Connecticut. Accident Analysis and Prevention 35, pp. 369-379.

Zhang, J., Lindsay, J., Clarke, K., Robbins, G., Mao, Y., 2000. Factors affecting the severity of motor vehicle traffic crashes involving elderly drivers in Ontario. Accident Analysis and Prevention 32, pp. 117-125.

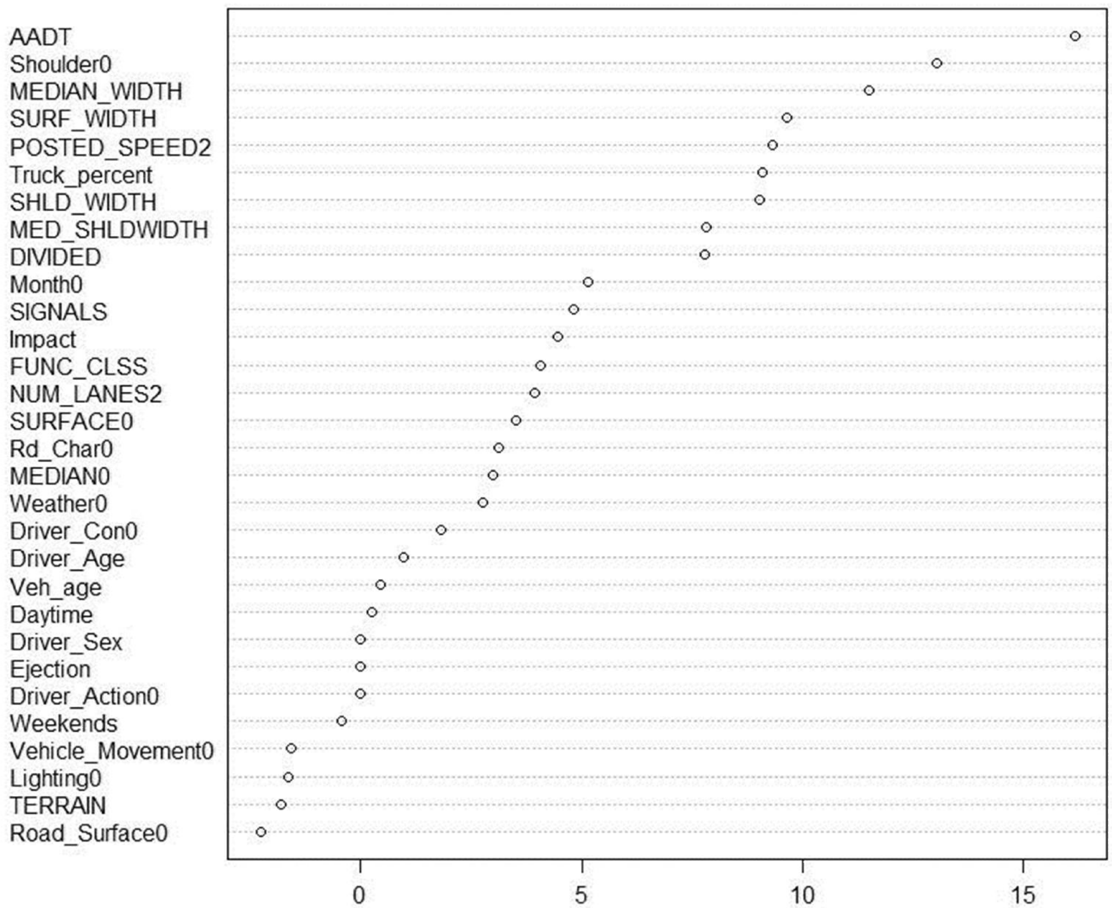
Zhu, X., Srinivasan, S., 2011. A comprehensive analysis of factors influencing the injury severity of large-truck crashes. Accident Analysis and Prevention 43, pp. 49-57.

Appendix A. Ranking of Important Variables in Random Forest



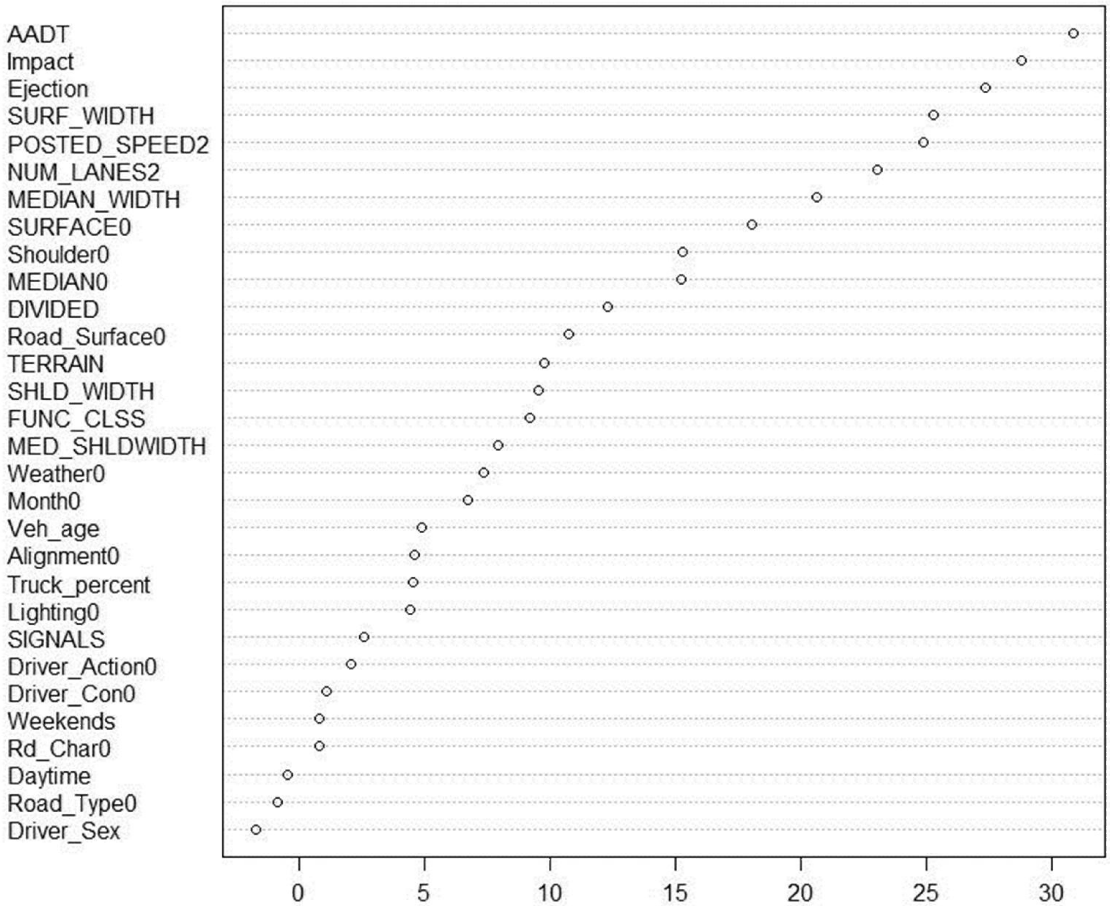
(a) C-H(C) crashes

Figure A-1. Rankings of Important Variables in Random Forests



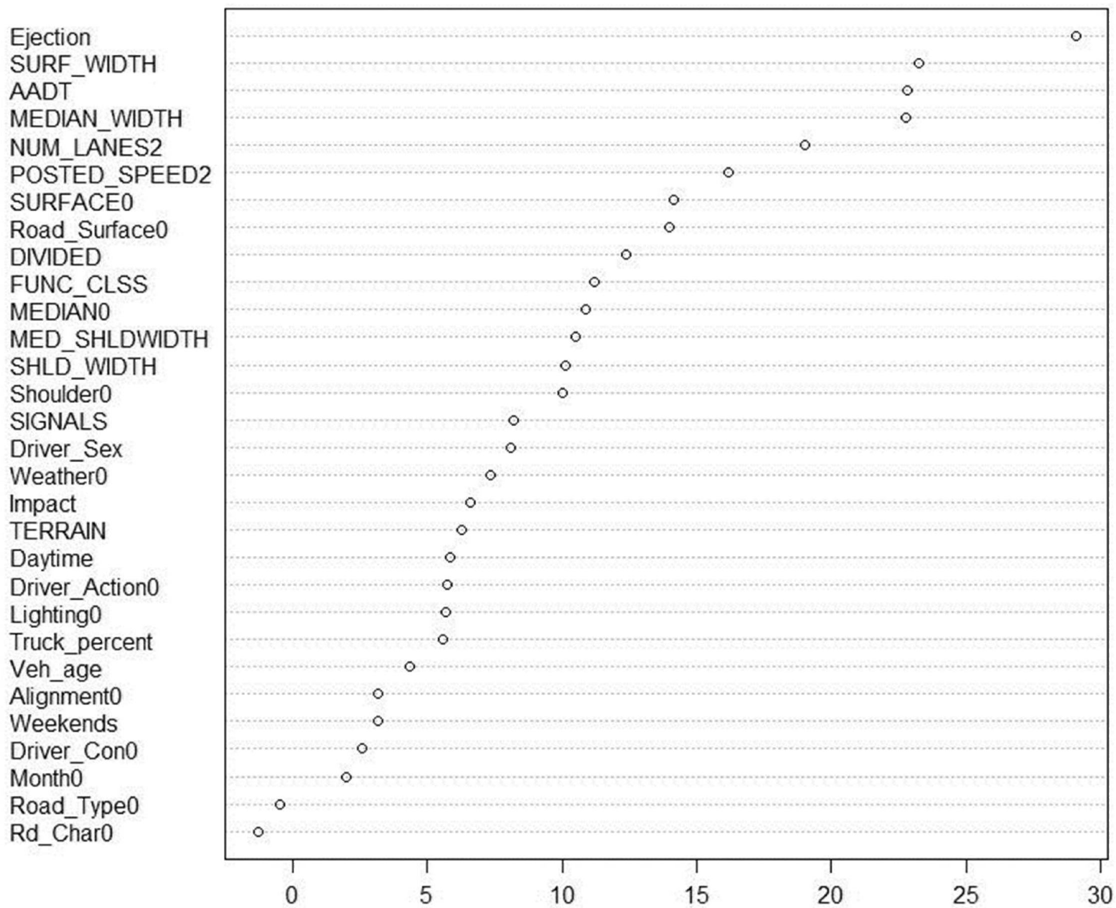
(b) C-H(H) crashes

Figure A-1. Rankings of Important Variables in Random Forests (Continued)



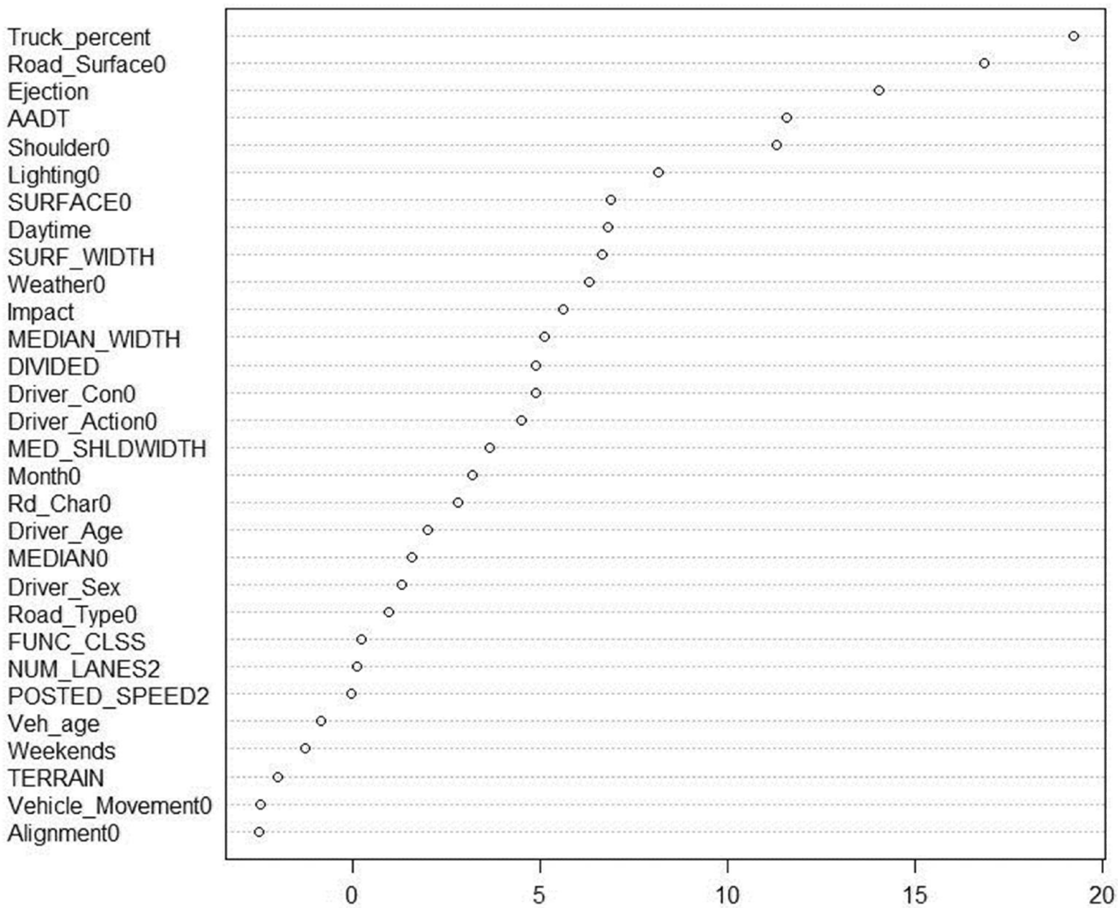
(c) C-L(C) crashes

Figure A-1. Rankings of Important Variables in Random Forests (Continued)



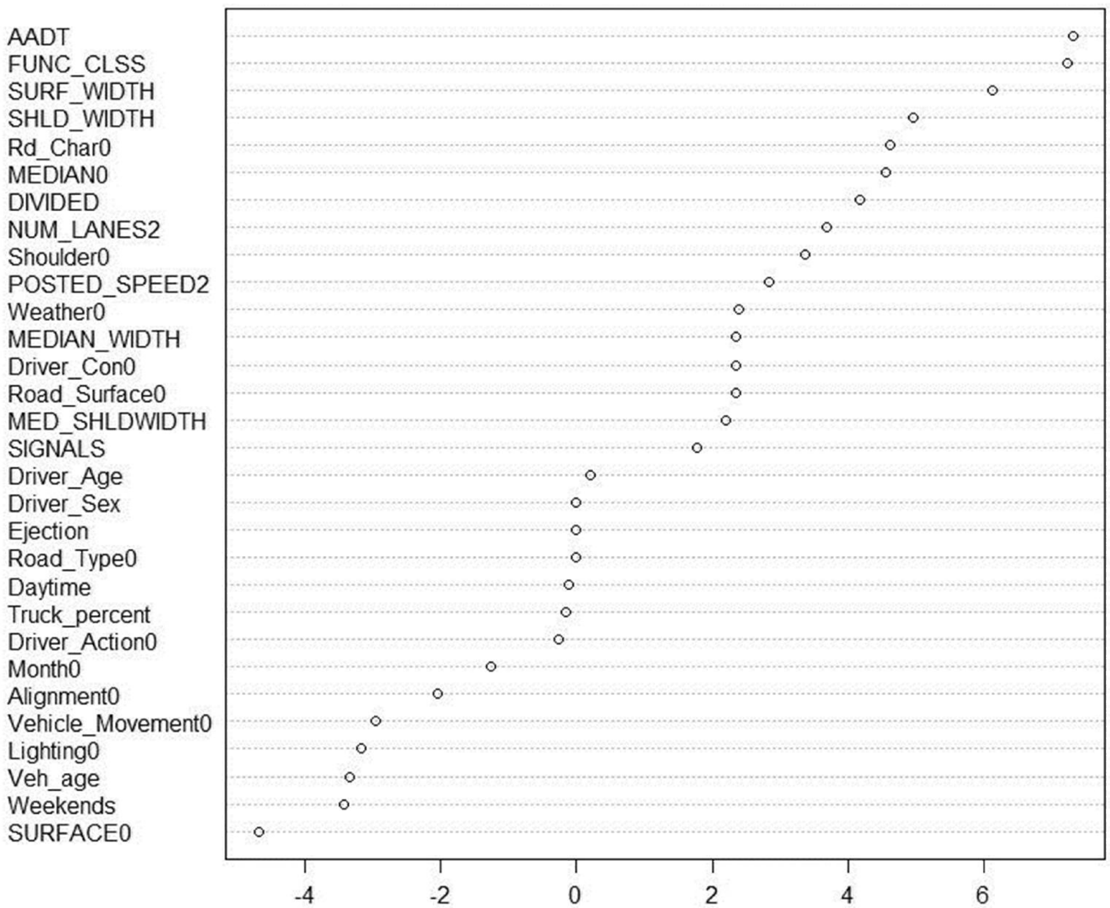
(d) C-L(L) crashes

Figure A-1. Rankings of Important Variables in Random Forests (Continued)



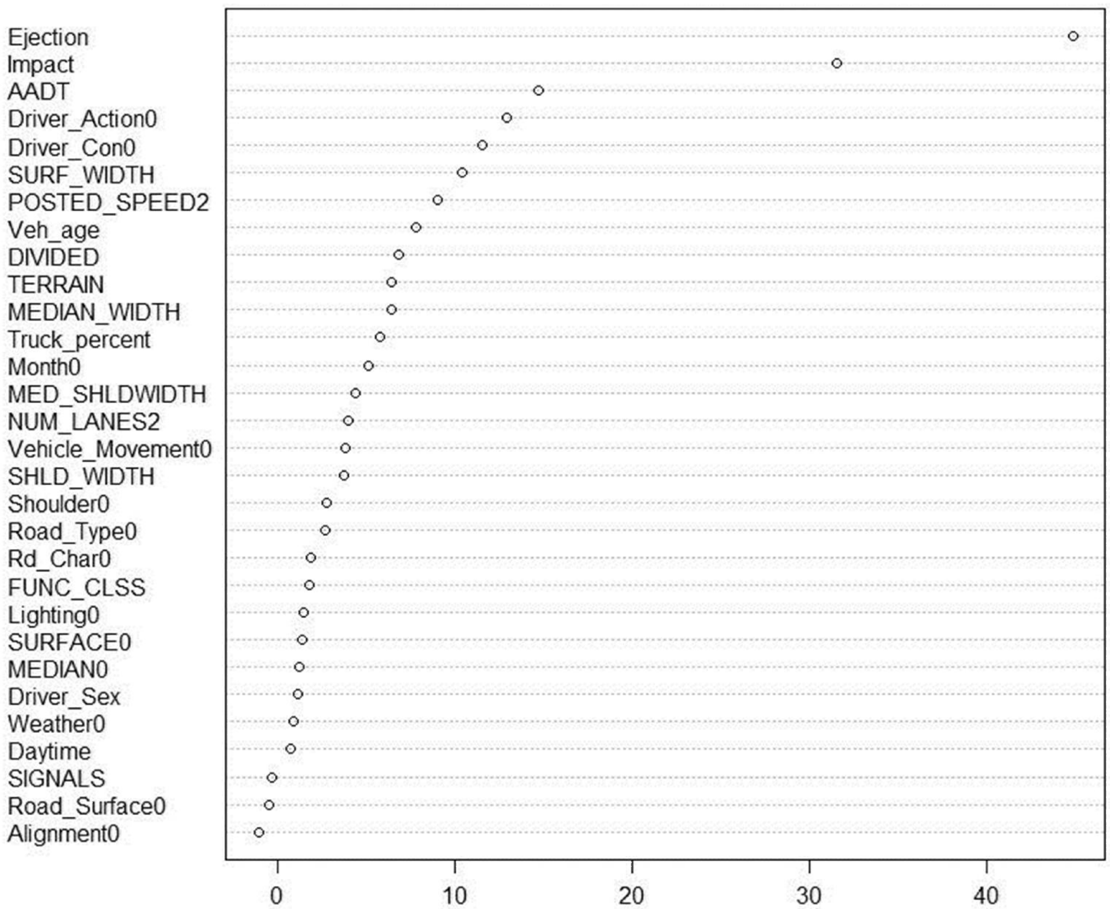
(e) H-H crashes

Figure A-1. Rankings of Important Variables in Random Forests (Continued)



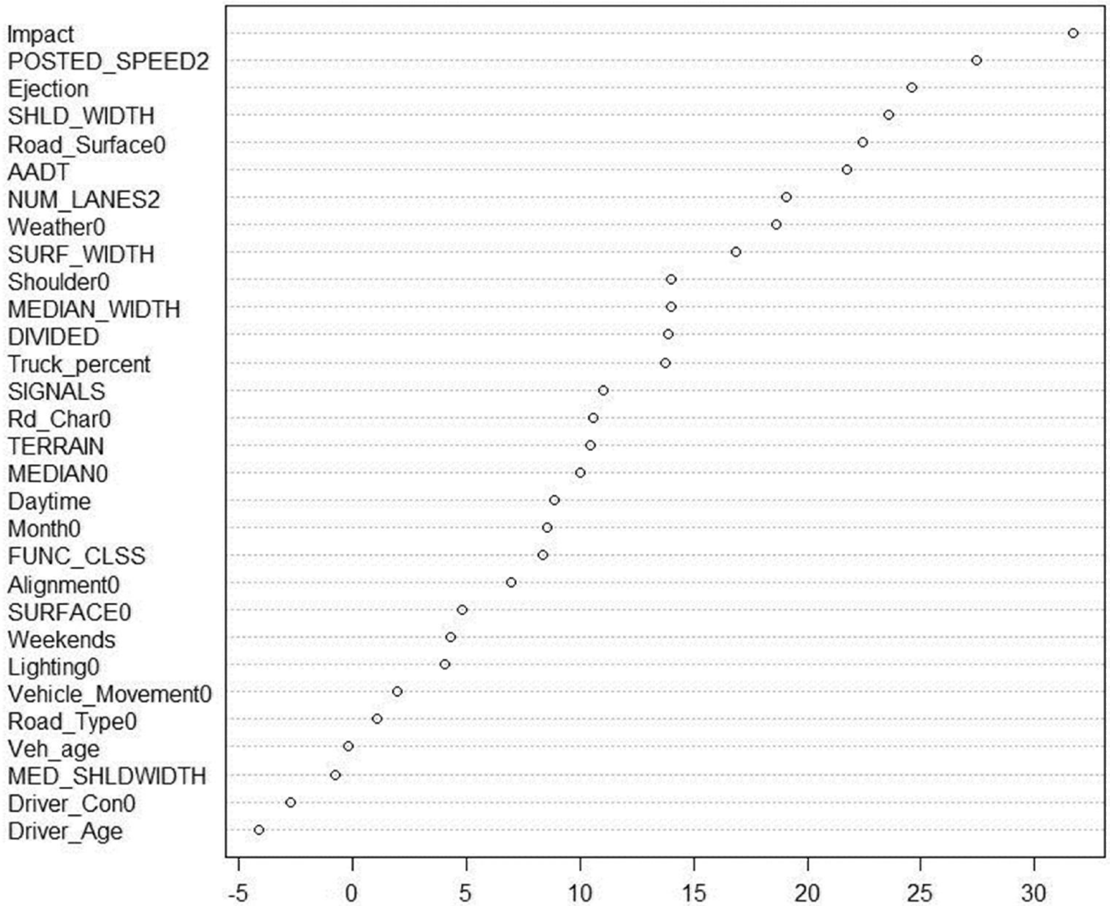
(f) L-H(H) crashes

Figure A-1. Rankings of Important Variables in Random Forests (Continued)



(g) L-H(L) crashes

Figure A-1. Rankings of Important Variables in Random Forests (Continued)



(h) L-L crashes

Figure A-1. Rankings of Important Variables in Random Forests (Continued)

Appendix B. Result of CART

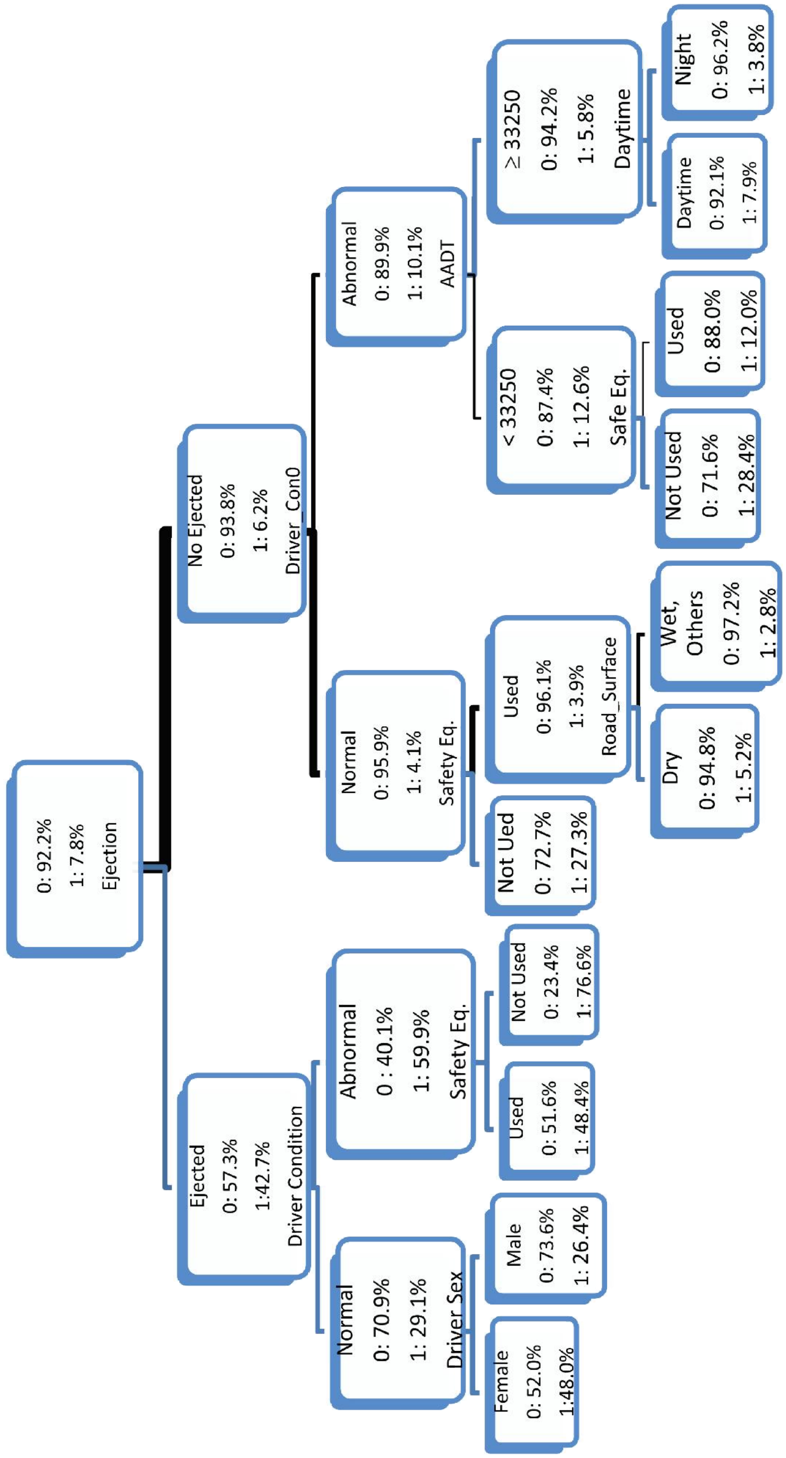


Figure B-1. CART for Single-Vehicle Crashes

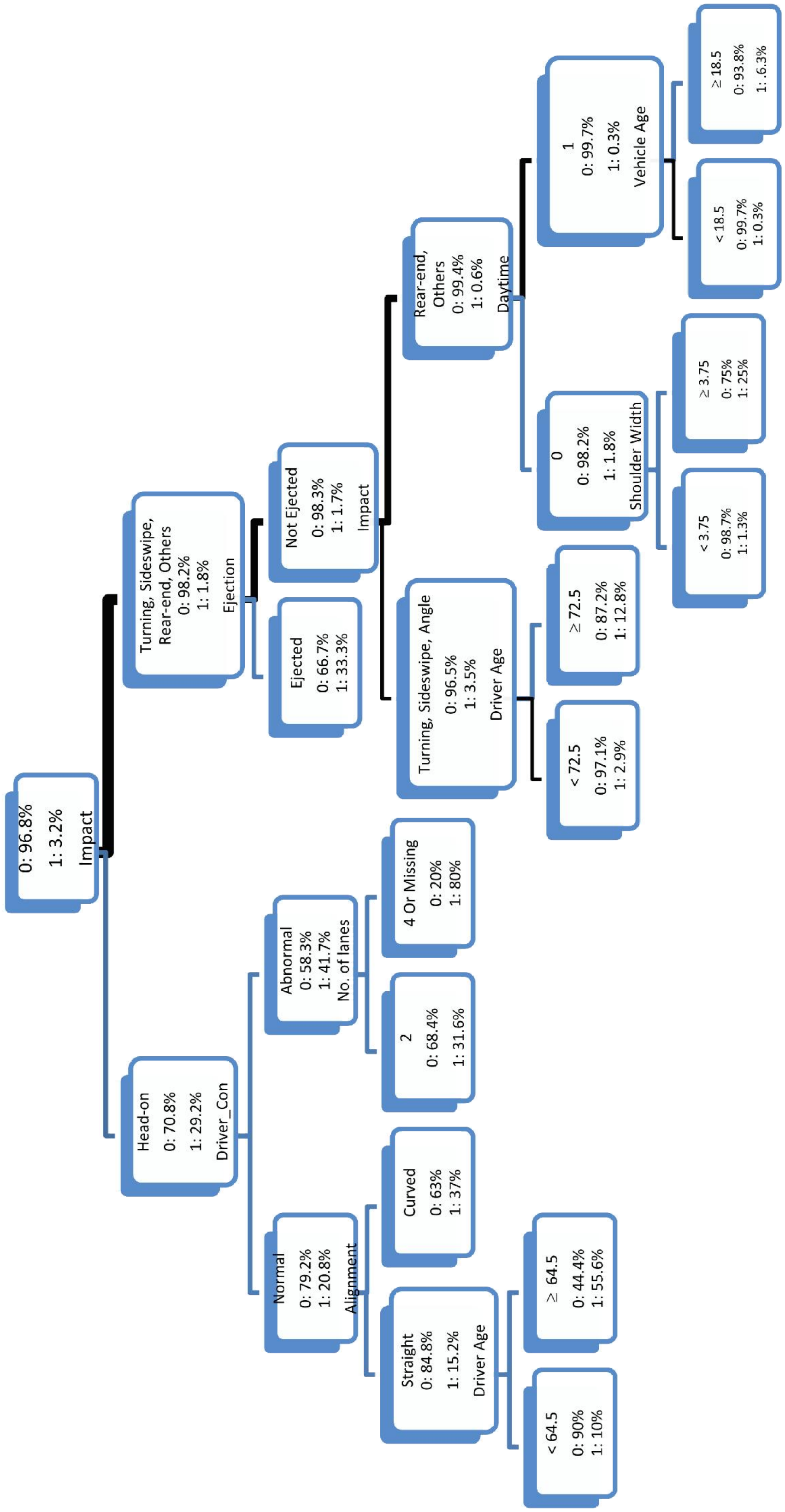


Figure B-2. CART for C-C Crashes

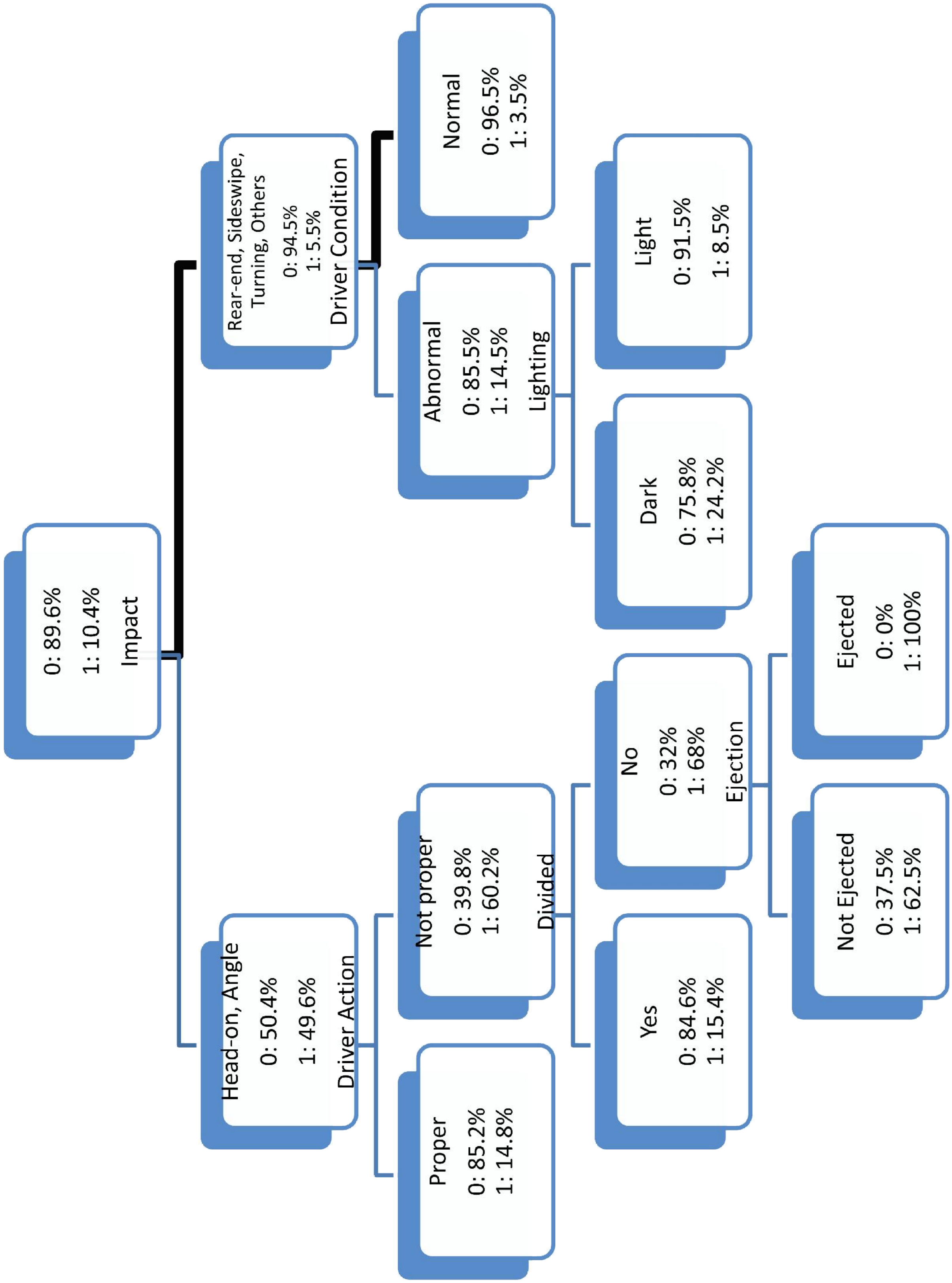


Figure B-3. CART for C-H(C) Crashes

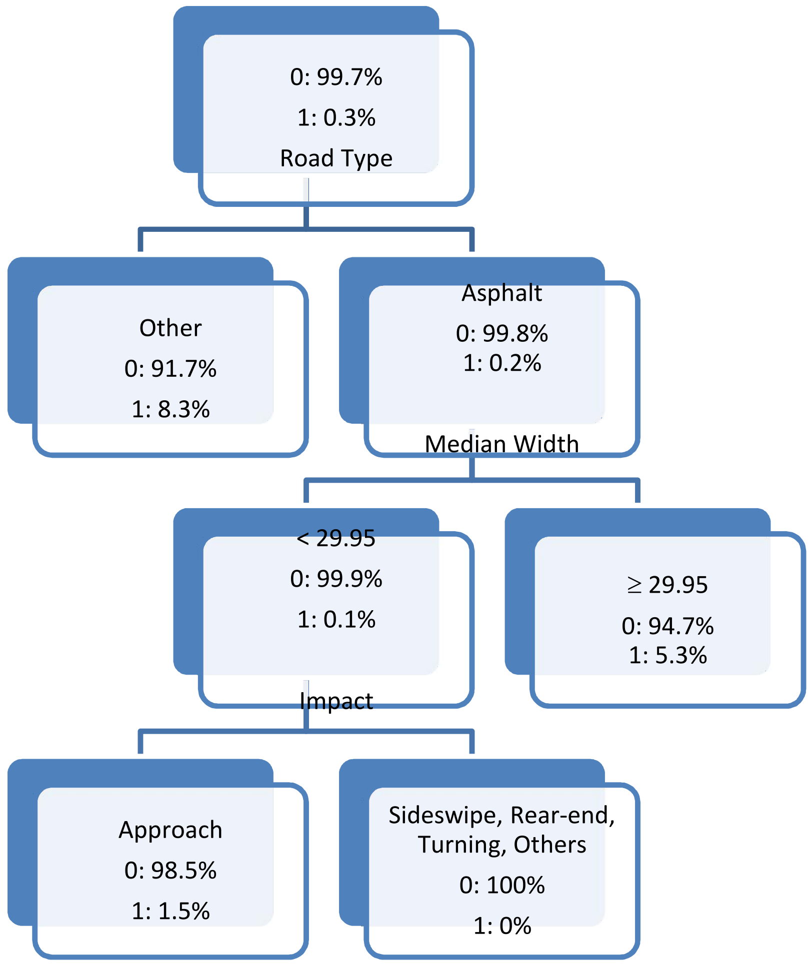


Figure B-4. CART for C-H(H) Crashes

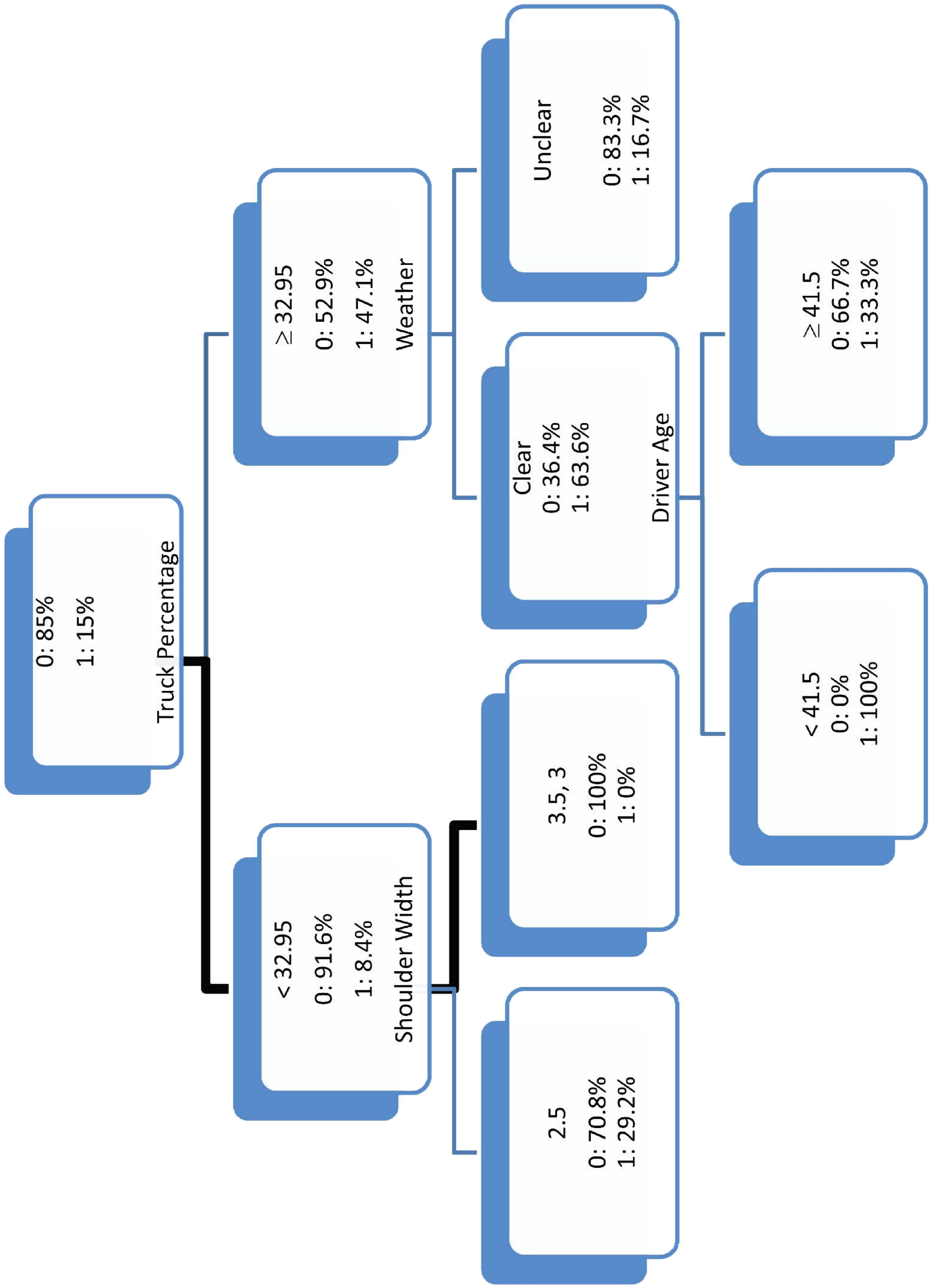


Figure B-5. CART for H-H Crashes

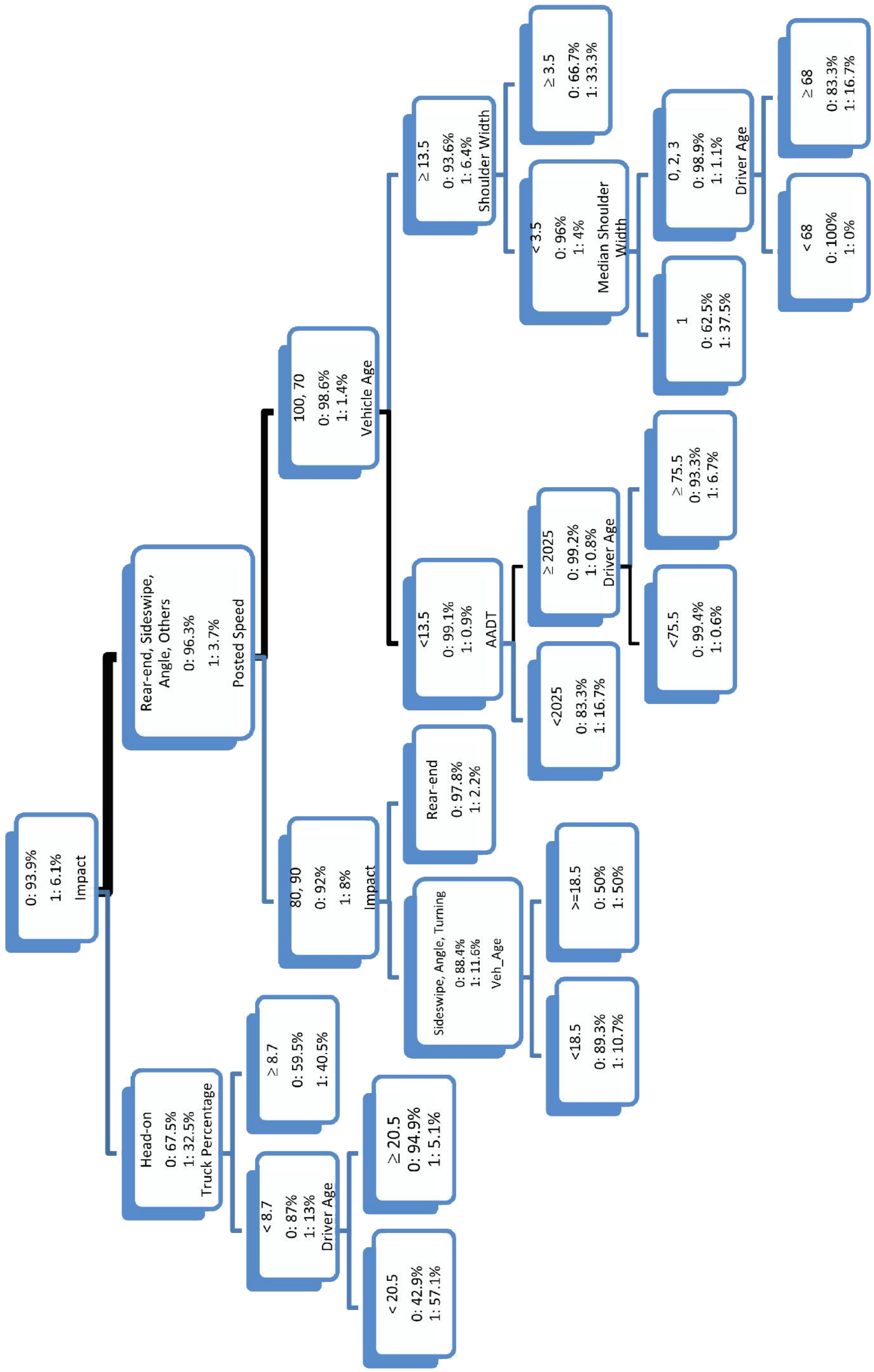


Figure B-6. CART for C-L(C) Crashes

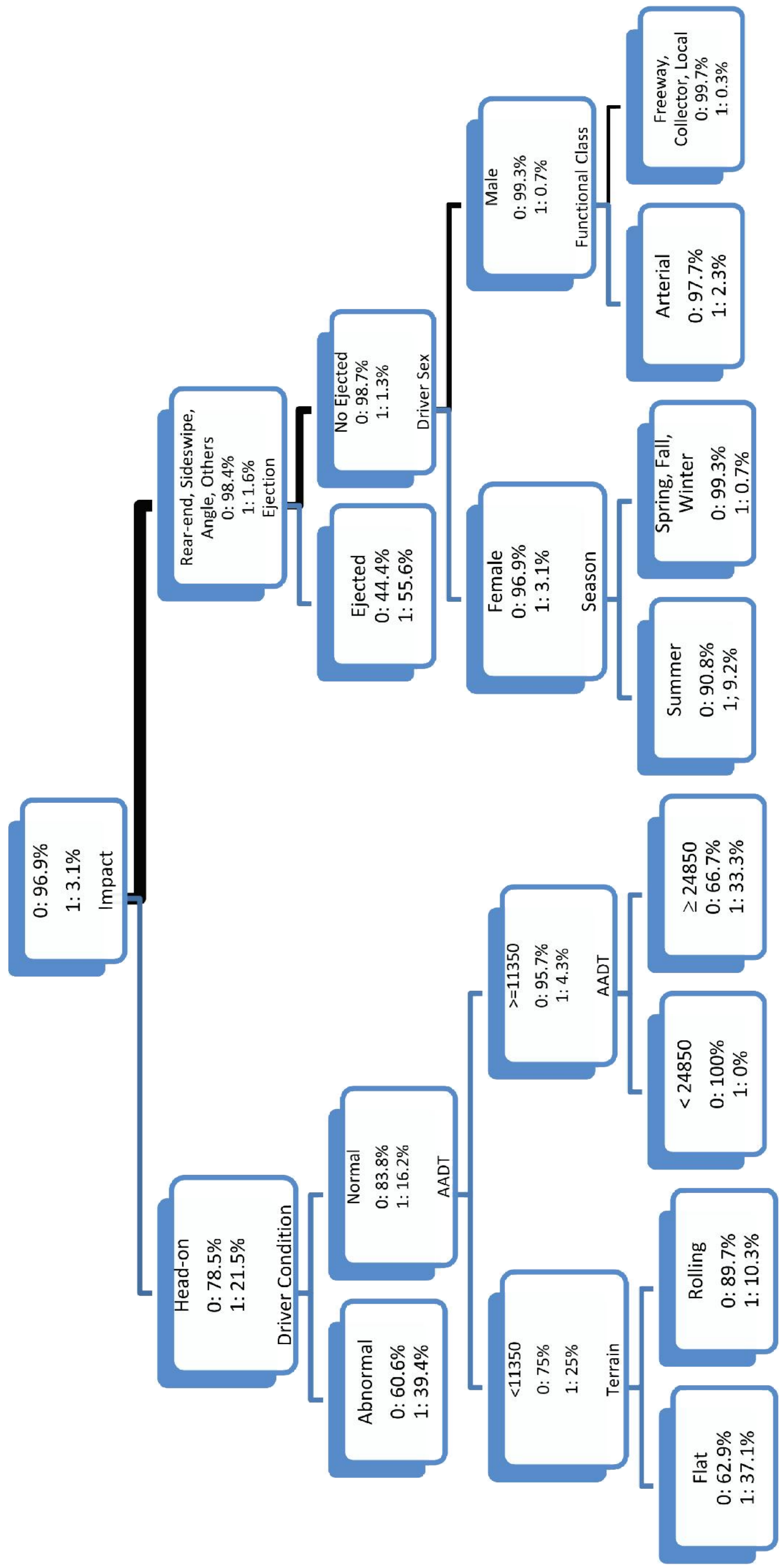


Figure B-7. CART for C-L(L) Crashes

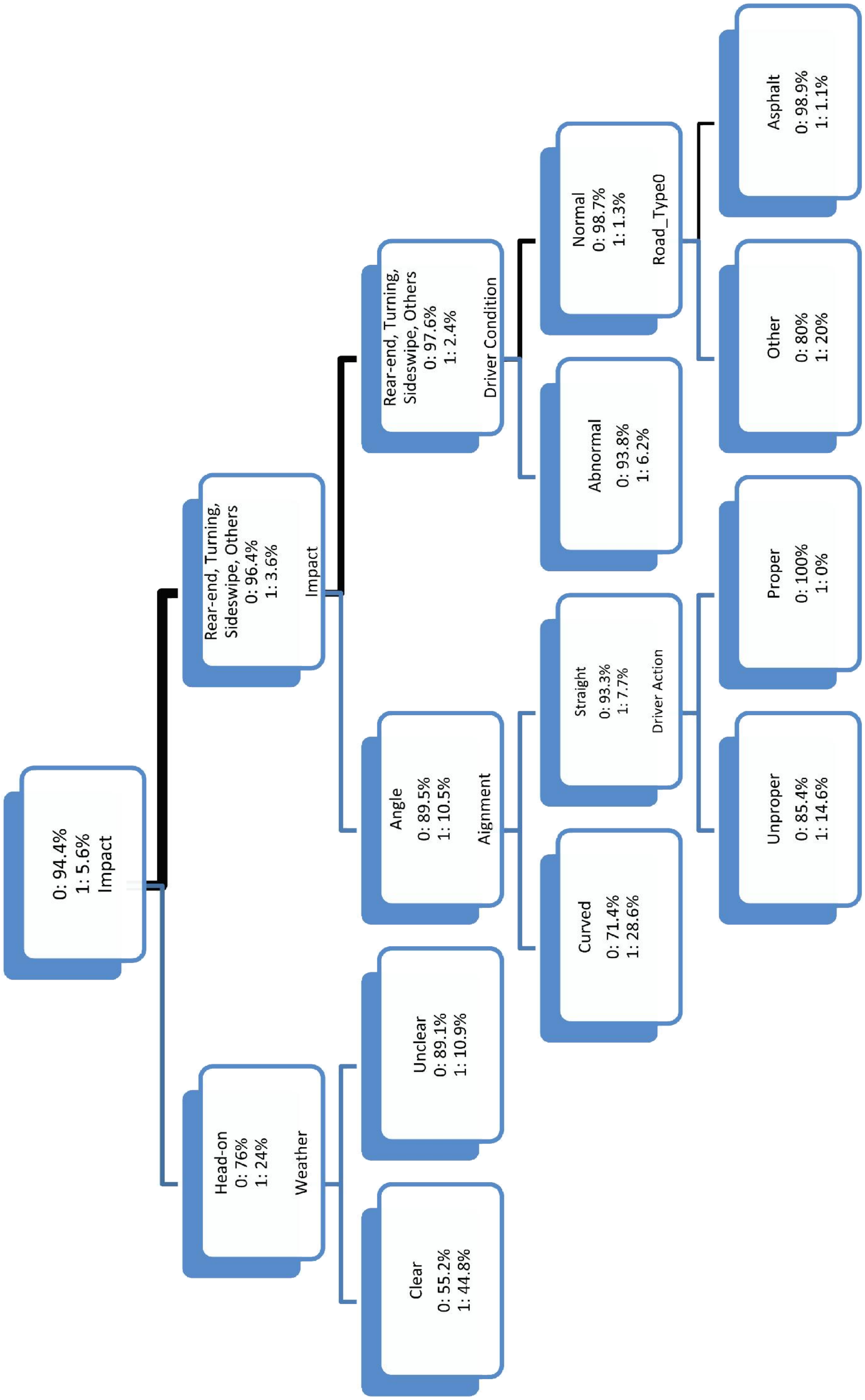


Figure B-8. CART for L-L Crashes

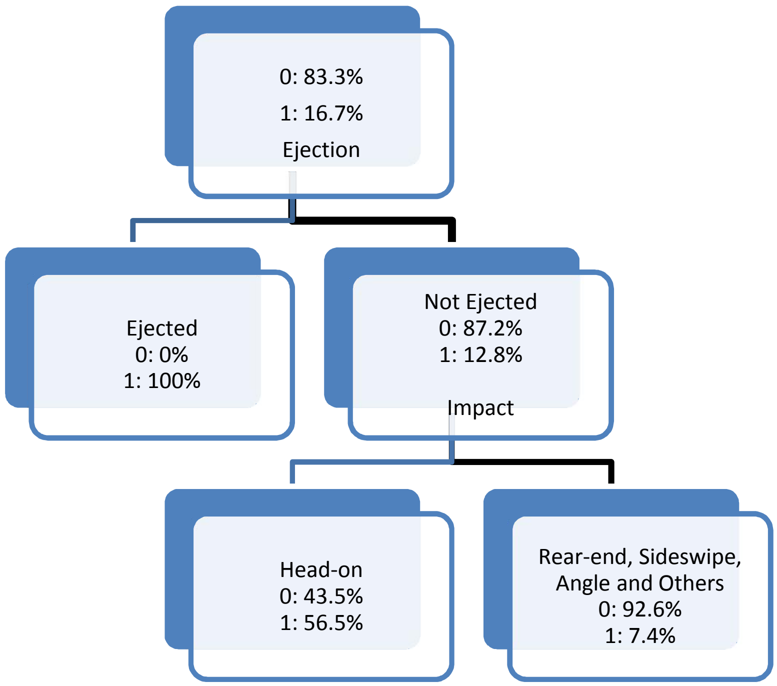


Figure B-9. CART for L-H(L) Crashes

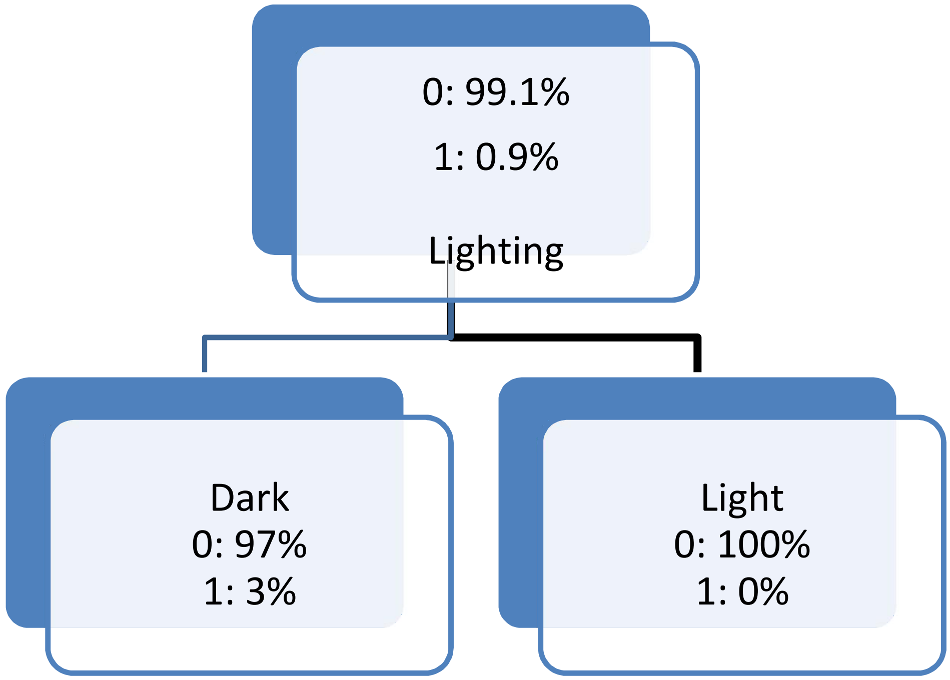


Figure B-10. CART for L-H(H) Crashes

Appendix C. Result of BRT

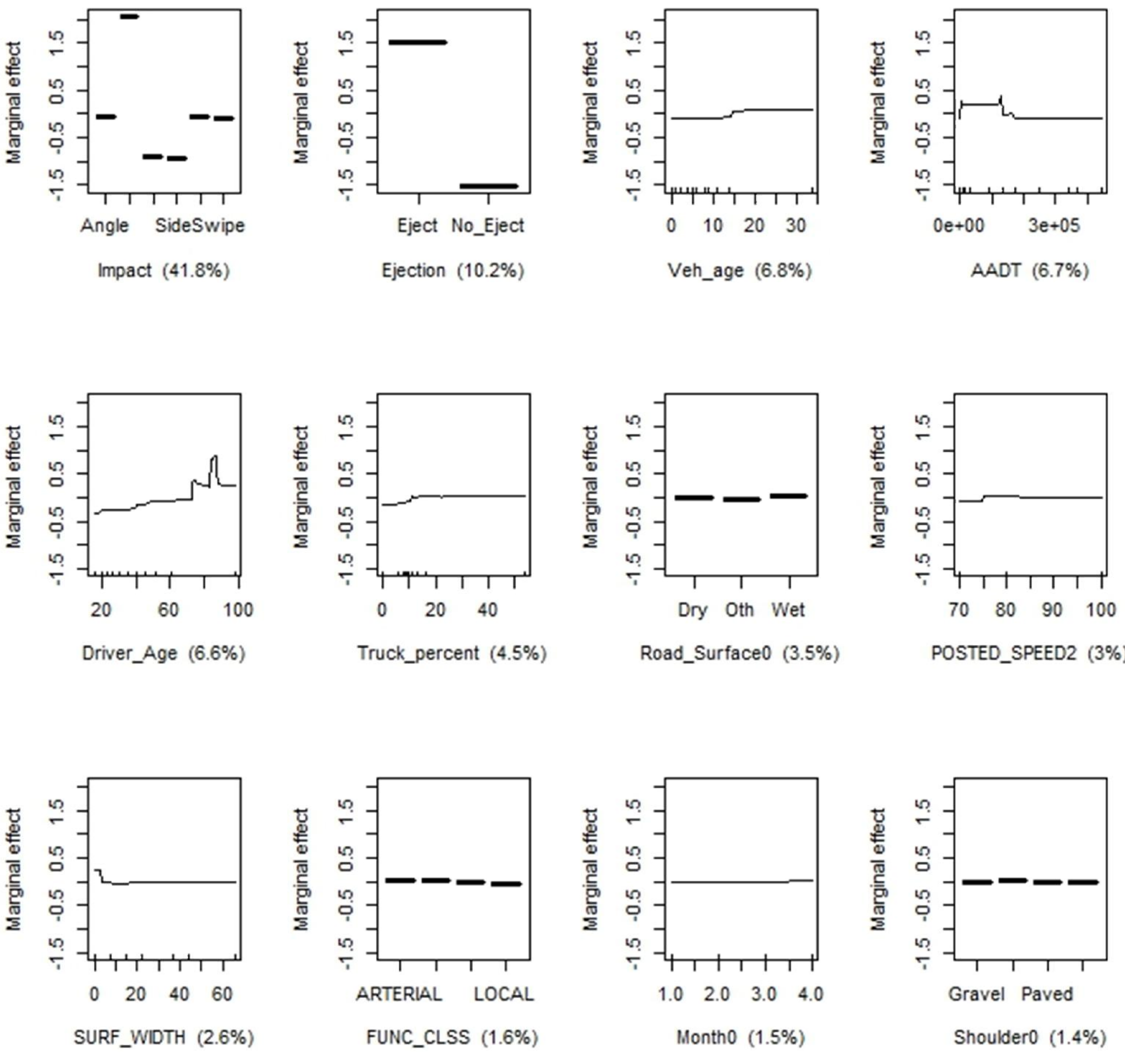


Figure C-1. Marginal Effects of the 12 Most Important Variables for C-C Crashes in BRT

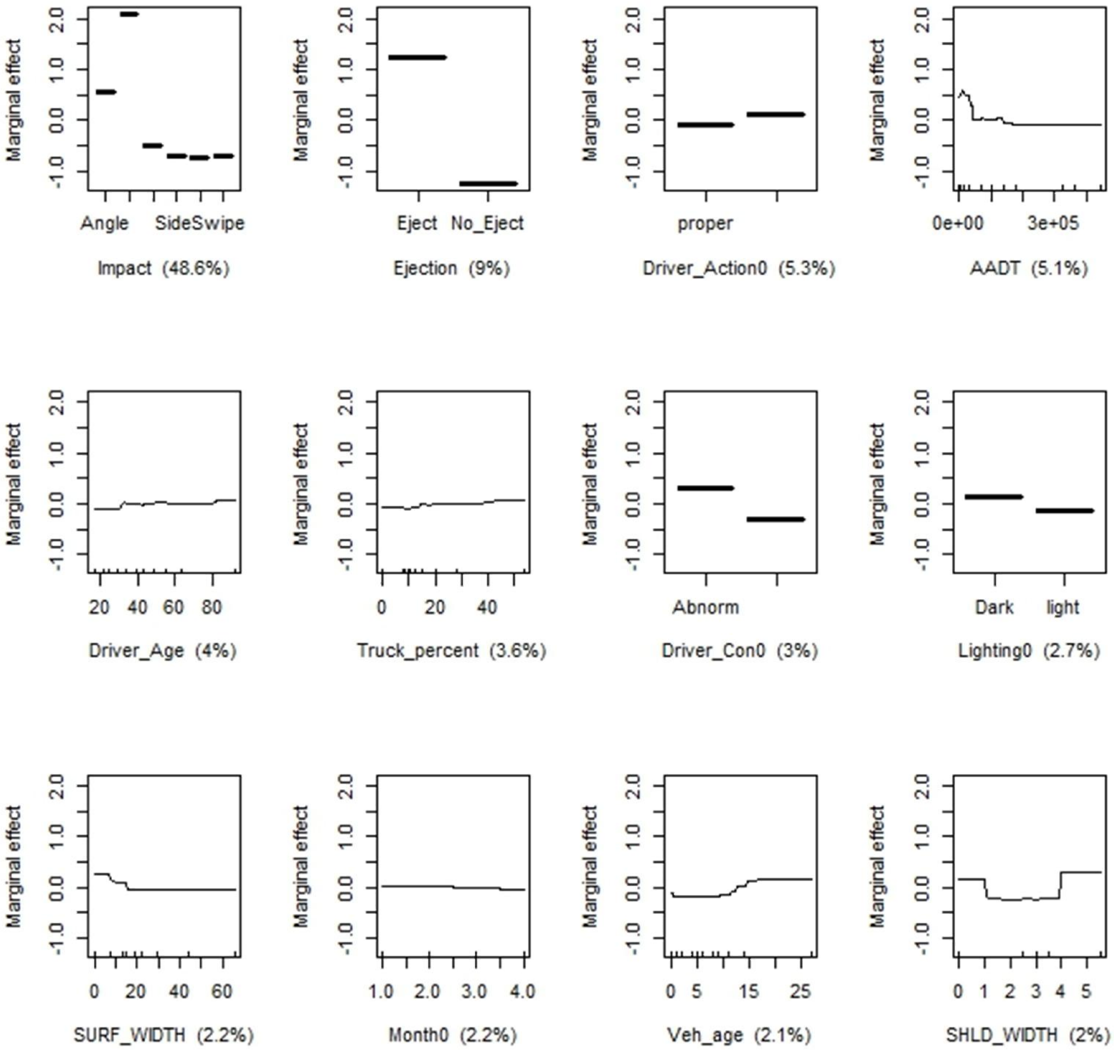


Figure C-2. Marginal Effects of the 12 Most Important Variables for C-H(C) Crashes in BRT

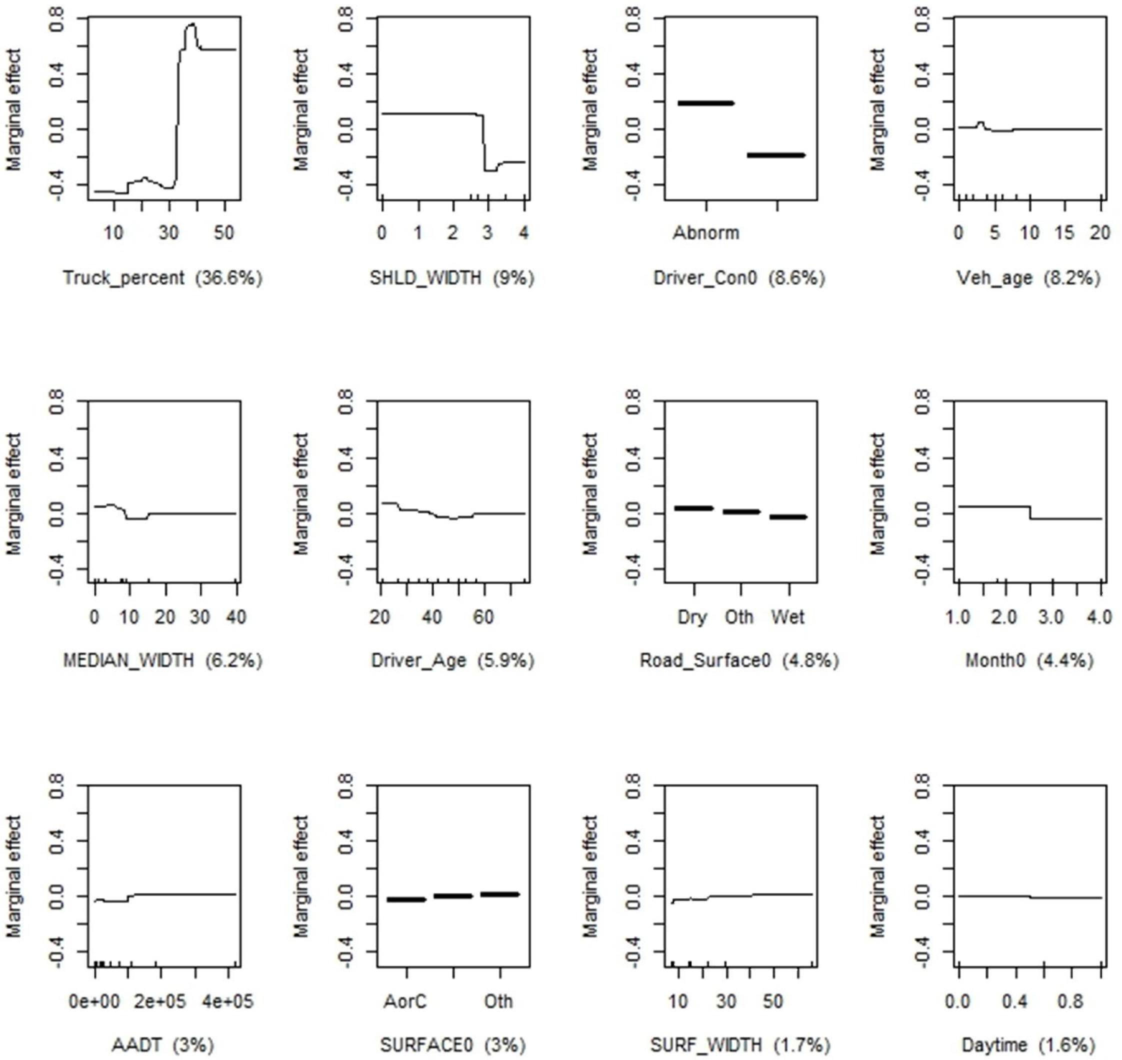


Figure C-3. Marginal Effects of the 12 Most Important Variables for H-H Crashes in BRT

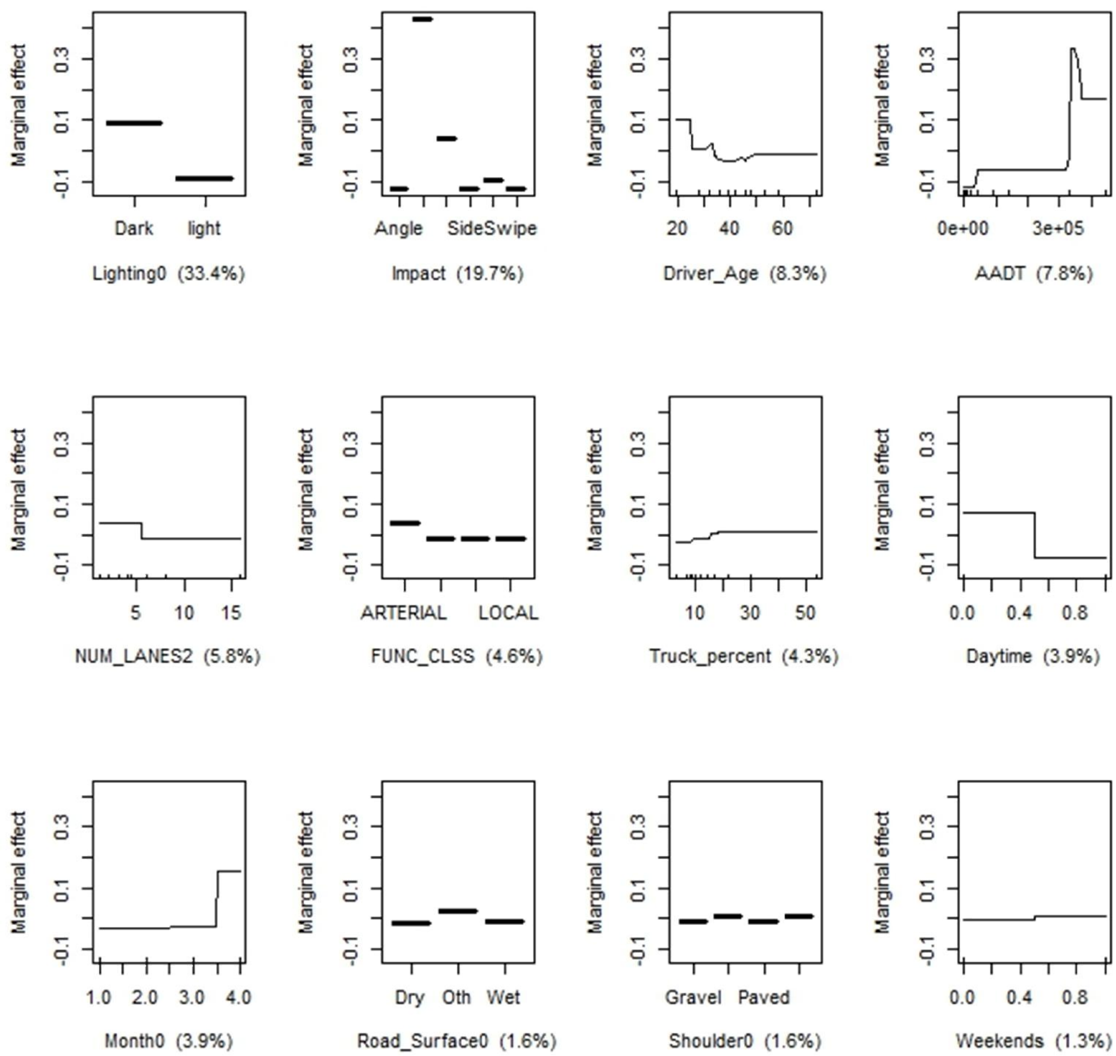


Figure C-4. Marginal Effects of the 12 Most Important Variables for L-L Crashes in BRT

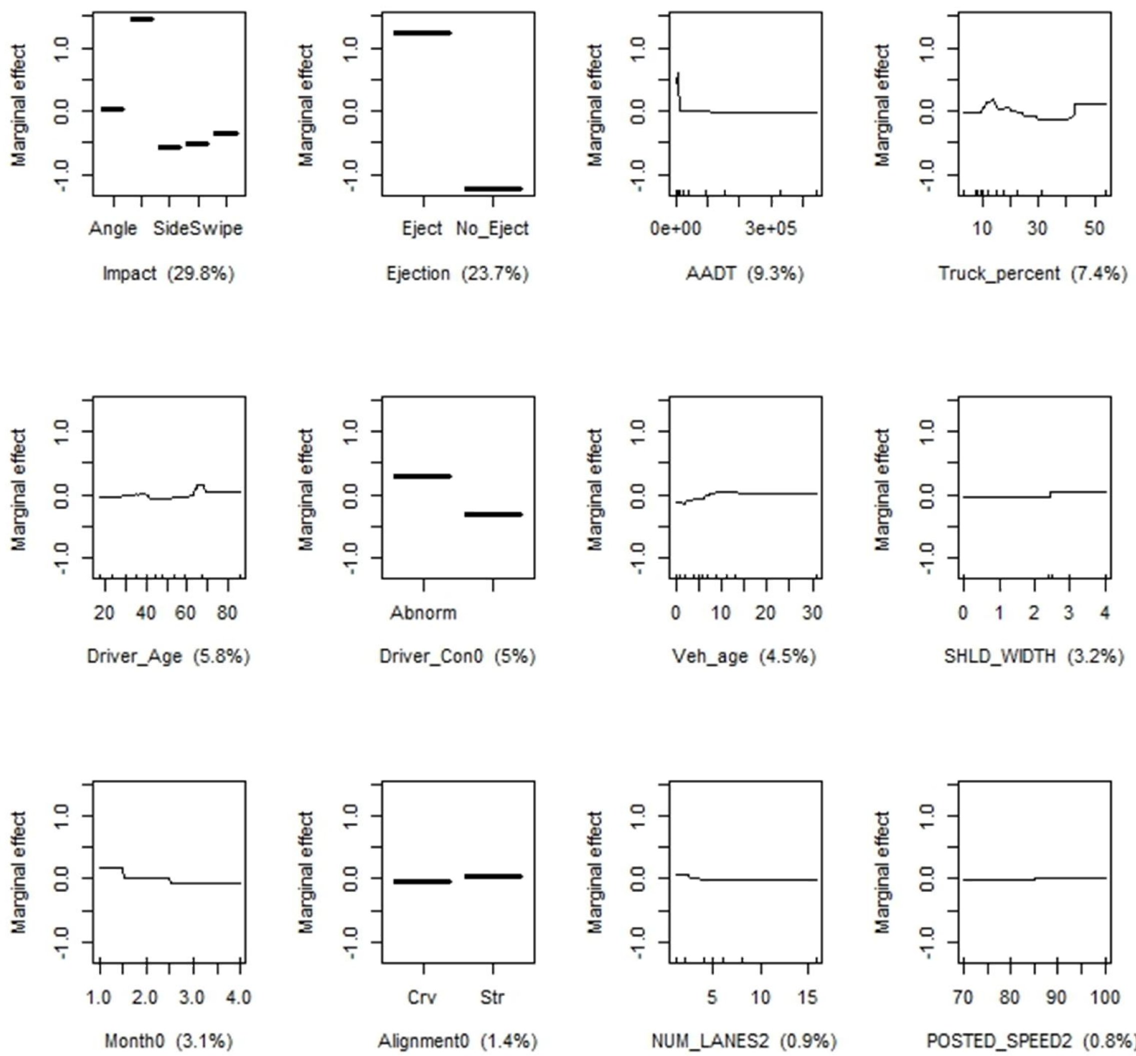


Figure C-5. Marginal Effects of the 12 Most Important Variables for L-H(L) Crashes in BRT

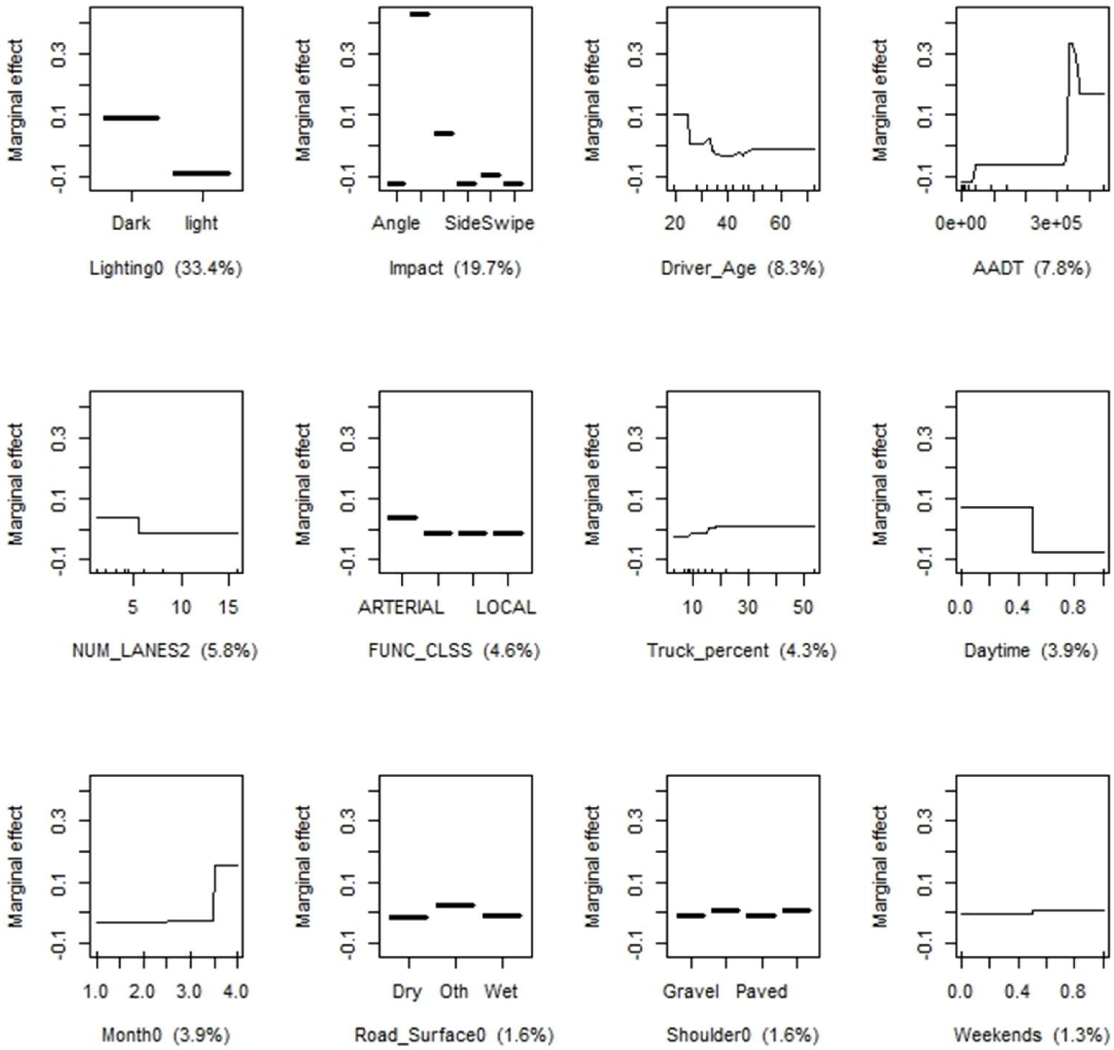


Figure C-6. Marginal Effects of the 12 Most Important Variables for L-H(H) Crashes in BRT

Appendix D. Result of HOL Models (2-Level Injury Severity)

In this section, the results of the HOL models with 2-level injury severity (severe or non-severe) were discussed. Table D-1 shows the result of the single-vehicle crash model. The effects of the variables on injury severity were generally similar to the HOL models with 4-level injury severity (Table 5-1).

Table D-1. Parameters of HOL Model for Single-Vehicle Crashes

Parameter	Estimate	Pr > t
Intercept	-3.72	<.0001
Speed limit (km/h)	0.01	0.0113
Passenger car (1 = passenger car; 0 = otherwise)	0.88	<.0001
Light truck (1 = light truck; 0 = otherwise)	0.41	0.0403
Heavy truck (1 = heavy truck; 0 = otherwise)	0.48	0.0312
Young (1 = age ≤ 30; 0 = otherwise)	-0.21	0.0011
Female (1 = female; 0 = male)	-0.28	<.0001
Safety equipments (1 = with safety equip.; 0 = no safety equip.)	-1.30	<.0001
Ejection (1 = ejected from vehicle; 0 = not ejected from vehicle)	2.17	<.0001
Number of lanes	-0.08	<.0001
Curved road (1=curved; 0= straight)	0.21	0.0019
Variance		
Passenger car	-0.42	0.0105
Log likelihood at convergence ($L^*(\beta)$)	-3204	
Log likelihood ratio index (ρ^2)	0.12	
Number of observations	13277	

However, there were some differences in the results between the 2-level and 4-level models. For instance, a negative effect of young drivers (≤ 30) was observed in the 2-level model unlike the 4-level model. This indicates that young drivers are less likely to be severely injured compared to older drivers (> 30). This is consistent with the finding of the past studies (e.g. Weiss et al. 2014). Also, a positive effect of male drivers was observed in the 2-level model unlike the 4-level model.

It was observed that the variance in random effects for passenger car was significant at a 95% significance level. This is potentially because driver's injury severity can greatly vary even if they are in the same vehicle type (passenger car). For example, drivers in a sedan may sustain different injury compared to drivers in a four-wheel-drive SUV or a minivan. Thus, the HOL model is more advantageous in capturing heteroscedasticity for more variable responses.

Also, less significant variance in the effects of variables among drivers was captured by the 2-level model than the 4-level model. This is because when injury severity levels decrease from four to two, the variance of injury severities also decreases. Thus, it is harder to capture the variation in the effects among observations using the 2-level model.

Table D-2 shows the results of two-vehicle crash models. Similar to single-vehicle crashes, the effects of variables on driver's injury severity were generally the same as 4-level HOL models. Some new geometric and environmental factors (which were not significant in the 4-level models) were also significant for C-C crashes. Injury severity was higher on the road with flat terrain and collector, but lower on the road with narrower median shoulder width. This indicates some road geometric and functional characteristics have strong effects on injury severity.

It is worth to note that truck percentage was significant for C-L(C) crashes unlike the 4-level model. This indicates that car driver's injury severity increases as there are more trucks on the road. This is potentially because of higher variation in speed and more complex driving conditions with higher truck percentage in the traffic stream. Also, a significant variation in random effects of truck percentage for C-L(C) crashes indicates that higher truck percentage increases variability of car driver's injury severity.

Table D-2. Parameters of HOL Models for Two-Vehicle Crashes

Parameter	C-C Estimate	Pr > t	C-H(C) Estimate	Pr > t	C-H(H) Estimate	Pr > t	H-H Estimate	Pr > t
Intercept	-0.07	0.9656	-0.43	0.5088	-3.67	0.0017	0.53	0.5206
Speed limit	0.04	0.0053	-	-	-	-	-	-
No. of streams	-0.46	0.0134	-	-	-	-	-	-
Safety equip.	-3.48	<.0001	-1.92	0.0019	-3.16	0.006	-2.38	0.004
Young (≤ 30)	-0.71	0.0021	-	-	-	-	-	-
Ejected	3.27	<.0001	2.88	<0.0001	-	-	-	-
Head-on	2.58	<.0001	2.75	<0.0001	-	-	2.08	0.0078
Rear-end	-1.59	<.0001	-	-	-	-	-	-
Divided road	-2.35	0.0145	-	-	-	-	-	-
Flat terrain	-[a]	-	0.55	0.0373	-	-	-	-
Abnormal driver condition	-	-	0.86	<0.0001	-	-	1.27	0.0109
AADT	-	-	-0.64	<0.0001	-	-	-	-
Improper driver action	-	-	0.60	0.0173	-	-	-	-
Dark	-	-	0.81	<0.0001	-	-	-	-
Median shoulder width	-	-	-0.25	0.0182	-	-	-	-
Collector	-	-	0.63	0.0347	-	-	-	-
Curved road	-	-	-	-	2.11	0.0017	-	-
Summer	-	-	-	-	1.67	0.0328	1.07	0.0271
Variance								
Undivided	0.63	0.0279	-	-	-	-	-	-
$L^*(\beta)$	-563		-381		-39		-61	
ρ^2	0.28		0.35		0.15		0.17	
No. of obs.	5532		1757		1757		170	

[a] A variable is excluded due to statistically insignificance of the variable at a 90% significance level.

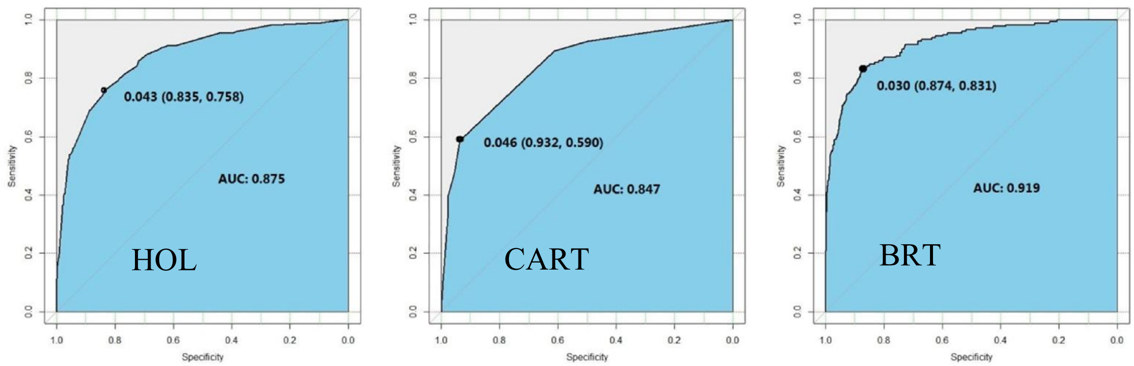
Table D-2. Parameters of HOL Models for Two-Vehicle Crashes (Continued)

	C-L(C)		C-L(L)		L-L	
Parameter	Estimate	Pr > t	Estimate	Pr > t	Estimate	Pr > t
Intercept	-3.21	0.0008	0.44	0.481	-2.66	<0.0001
Safety equip.	-2.00	<0.0001	-2.41	<0.0001	-1.56	0.0034
Rear-end	-1.09	<0.0001	-1.18	0.0008	-	-
AADT	-0.87	0.0003	-	-	-	-
Young (≤ 30)	-0.49	0.0109	-	-	-	-
Middle1 (31-45)	-0.57	0.007	-	-	-	-
Middle2 (46-60)	-0.49	0.021	-	-	-	-
Summer	-0.44	0.0212	-	-	-	-
Spring	-0.41	0.0275	-	-	2.13	<0.0001
Vehicle age	0.03	0.0226	-	-	-	-
Speed limit	0.03	0.001	-	-	-	-
Truck percentage	0.03	<0.0001	-	-	-	-
Undivided road	0.74	0.0023	-	-	-	-
Head-on	1.24	<0.0001	1.51	<0.0001	1.90	0.0019
Ejected	1.45	0.0059	2.77	<0.0001	1.33	0.0125
Angle	-[a]	-	-	-	0.85	0.0139
Abnormal driver condition	-	-	0.61	0.0178	-	-
Fall	-	-	-0.68	0.0292	-	-
Female	-	-	0.69	0.0038	-	-
Paved shoulder	-	-	-1.06	0.0011	-	-
Sideswipe	-	-	-	-	0.56	0.0625
Wet surface	-	-	-	-	0.64	0.0311
Turning	-	-	-1.01	0.0159	-	-
Variance						
Truck percentage	-0.03	0.0193	-	-	-	-
Clear weather	-	-	-	-	-1.60	0.0154
$L^*(\beta)$	-514		-324		-191	
ρ^2	0.28		0.25		0.30	
No. of obs.	3132		3128		1277	

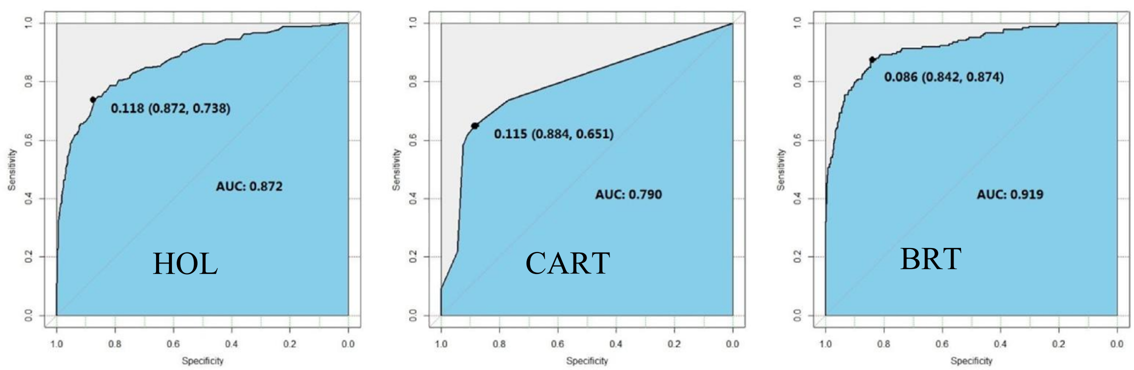
	L-H(L)		L-H(H)	
Parameter	Estimate	Pr > t	Estimate	Pr > t
Intercept	2.91	0.0074	-5.09	<.0001
Head-on	2.56	<.0001	2.47	<.0047
Ejected	3.86	<.0001	-	-
Abnormal driver condition	1.07	0.0049	-	-
Safety equipment	-2.58	0.0002	-	-
Middle2 (46-60)	-0.98	0.0429	-	-
Arterial	-1.73	0.0476	-	-
Collector	-2.51	0.0105	-	-
Freeway	-2.14	0.0145	-	-
Variance				
Ejected	-5.41	<0.0001	-	-
$L^*(\beta)$	-102		-23	
ρ^2	0.40		0.13	
No. of obs.	372		370	

[a] A variable is excluded due to statistically insignificance of the variable at a 90% confidence level.

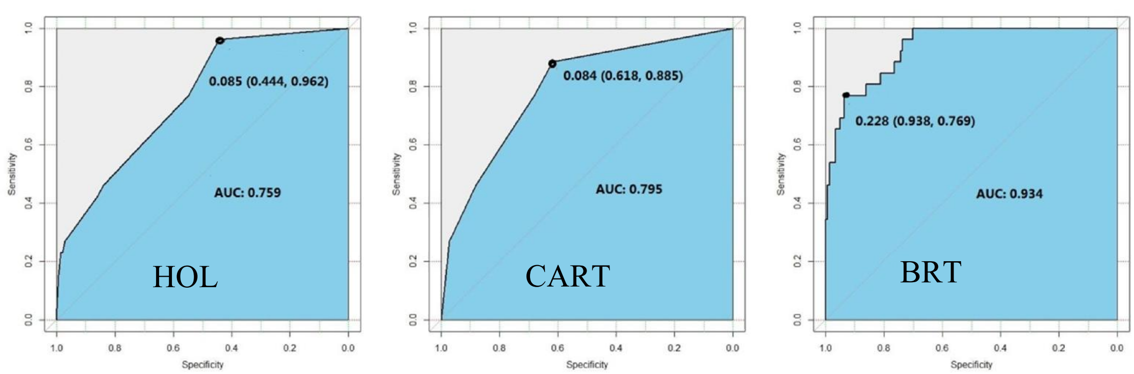
Appendix E. Comparison of ROC Curves



(a) C-C crashes

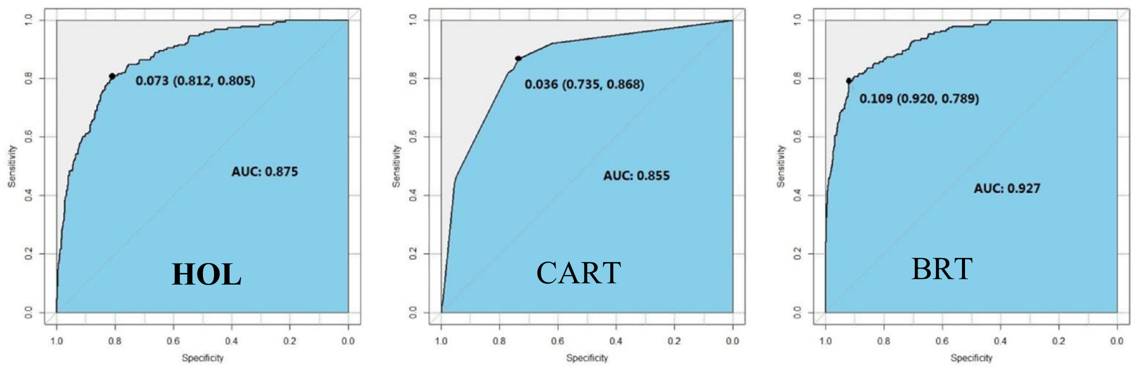


(b) C-H(C) crashes

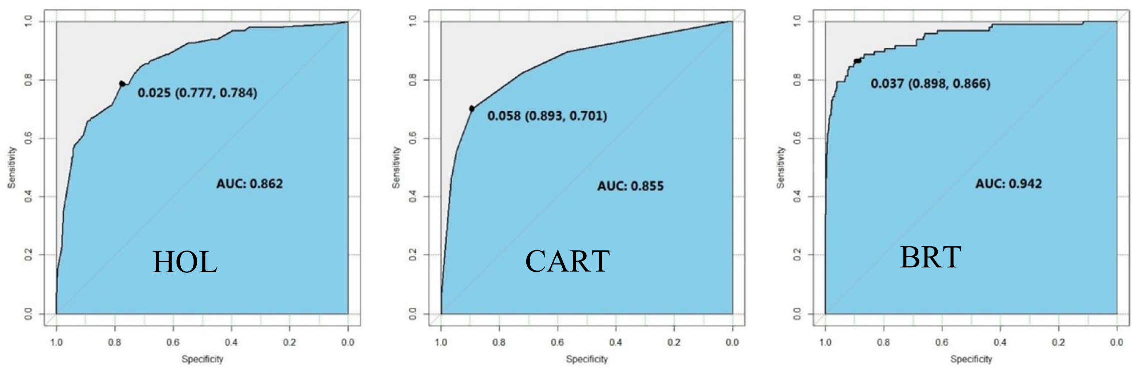


(c) H-H crashes

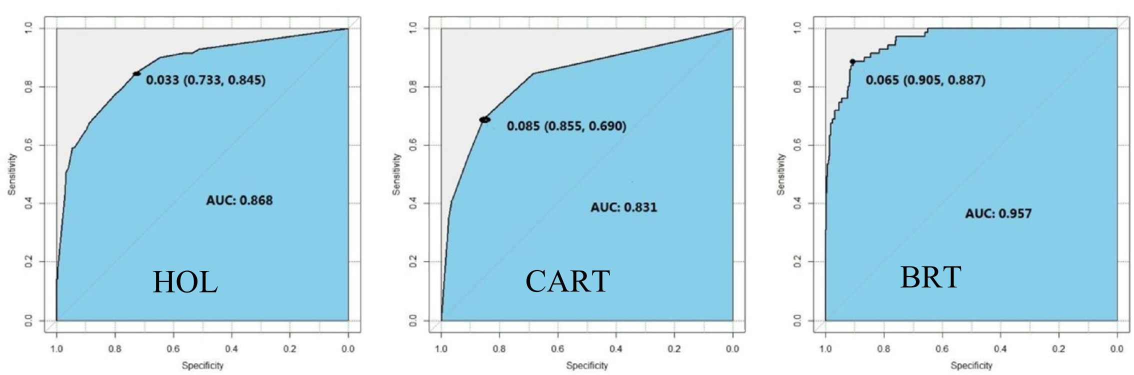
Figure E-1. Comparison of Goodness-of-fit Among HOL, CART and BRT Models Using ROC Curves



(d) C-L(C) crashes

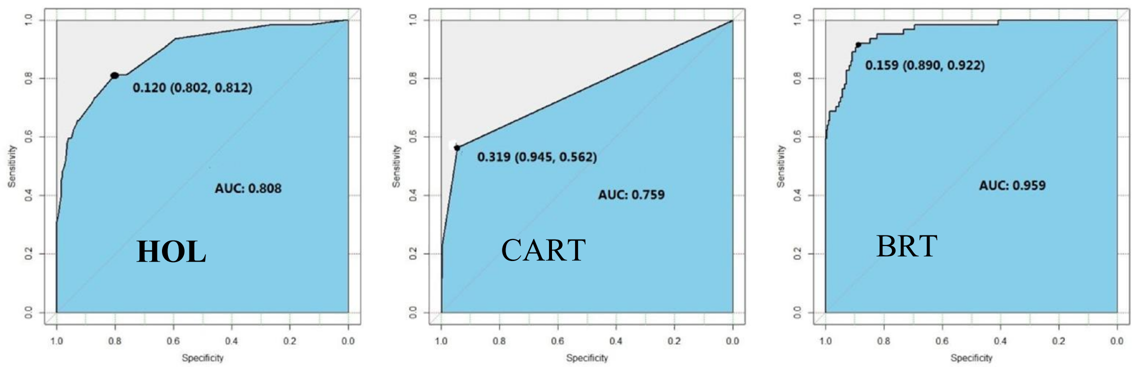


(e) C-L(L) crashes

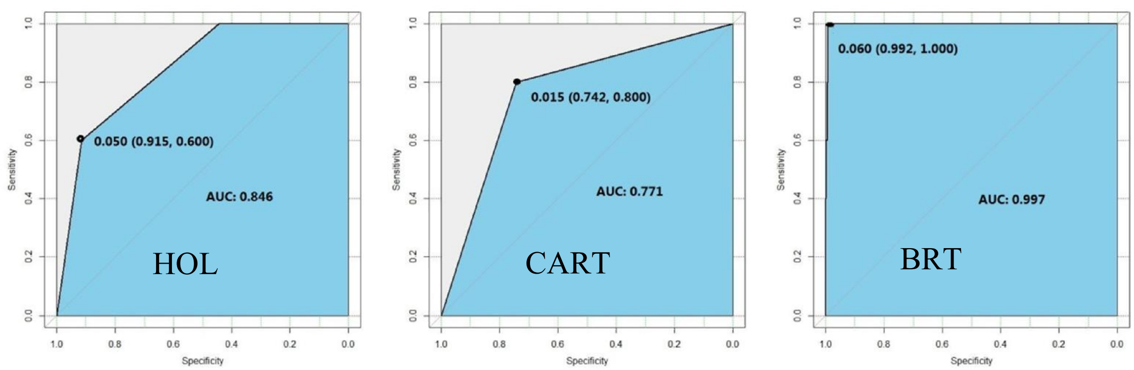


(f) L-L crashes

Figure E-1. Comparison of Goodness-of-fit among HOL, CART and BRT Models using ROC Curves (Continued)



(g) L-H(L) crashes



(h) L-H(H) crashes

Figure E-1. Comparison of Goodness-of-fit among HOL, CART and BRT Models using ROC Curves (Continued)

Appendix F. Copyright Permissions

Chris Lee 12:11 PM (6 minutes ago)

to li132

Dear Xuancheng,

I grant you permission to use my copyright work in your thesis.

Best regards,

Chris Lee, Ph.D., P.Eng.
Associate Professor
Department of Civil and Environmental Engineering
University of Windsor
401 Sunset Ave.
Windsor, ON N9B 3P4
Tel: (519) 253-3000 Ext. 2544
Fax: (519) 971-3686

-----Xuancheng Li <li132@uwindsor.ca> wrote: -----
To: Chris Lee <cclee@uwindsor.ca>
From: Xuancheng Li <li132@uwindsor.ca>
Date: 09/10/2014 11:47AM
Subject: Copyright Permissions

Hi, Dr.lee

I would like to ask your permissions of using some of the previous study results in the following published or in press papers:

Lee, C., Li, X., 2014. Analysis of Injury Severity of Drivers Involved in Single- and Two-Vehicle Crashes on Highways in Ontario. Accident Analysis and Prevention, Vol. 71, pp. 286-295.

Lee, C., Li, X., 2014. Predicting Driver's Severe Injury in Single-Vehicle and Two-Vehicle Crashes Using Boosted Regression Trees. Submitted for presentation at 94th Transportation Research Board Annual Meeting and publication in Transportation Research Record: Journal of the Transportation Research Board.

I would like to use HOL model results and CART and BRT model results in my major thesis part five (results and discussion chapter) for driver injury severity analysis. I would like to discuss these results and draw some conclusions for completing my thesis.

Could you please confirm your consent to this proposed use of your copyright works by return of email.

I look forward to hearing from you.

With kind regards

Xuancheng Li

M.A.Sc. in Civil Engineering candidate

Department of Civil and Environmental Engineering

University of windsor

VITA AUCTORIS

NAME:	Xuancheng Li
PLACE OF BIRTH:	Shijiazhuang, Hebei, P.R. China
YEAR OF BIRTH:	1986
EDUCATION:	China University of Geoscience Great wall College Baoding, Hebei, China B.Sc. in Civil Engineeing 2005-2009
	University of Windsor Windsor, Ontario, Canada M.A.Sc. in Civil Engineering 2012-2014